Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Statistical Theory and Methods for Comparing Distributions

Presented by: Ilmun Kim

Accepted by: Department of Statistics & Data Science

Readers:

_____
Larry Wasserman, Co-Chair

_____
Sivaraman Balakrishnan, Co-Chair

_____
Arthur Gretton

_____
Aaditya Ramdas

_____
Alessandro Rinaldo

Approved by the Committee on Graduate Degrees:

_____
Richard Scheines, Dean Date

Carnegie Mellon University

Statistical Theory and Methods for Comparing Distributions

A Dissertation Submitted to the Graduate School

in Partial Fulfillment of the Requirements for the degree

Doctor of Philosophy

in

Statistics

by

Ilmun Kim

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University

May 2020

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my advisors, Larry Wasserman and Sivaraman Balakrishnan, for their constant guidance and unwavering support. Throughout my PhD studies, they shaped my research path by giving me an ample amount of freedom to pursue my academic interests and steering me in the right direction when I was struggling. They were a constant source of encouragement and inspiration at every step of this long journey. Without their patience and invaluable advice, this dissertation would not have been possible.

I would like to extend my sincere thanks to Arthur Gretton, Aaditya Ramdas and Alessandro Rinaldo for being on my committee and providing me with valuable feedback. Their papers and research have certainly influenced my perspectives on research and have inspired many chapters in this dissertation. Special thanks to Aaditya who kindly invited me onto projects and taught me how to write a great research paper. I gratefully acknowledge the effort of Alessandro who has made our PhD program more enjoyable and taught me various topics in high-dimensional statistics through his classes. I am also deeply indebted to Ann Lee and Peter Freeman for their invaluable advice, guidance and support over my ADA project. In particular, Ann has helped engage me in various collaborative problems. I would also like to acknowledge Jing Lei and Matey Neykov for their collaborative efforts and for sharing their expertise with me. I am thankful to Yuting Wei for her advice on job applications and for her inspiring lectures. I benefited tremendously from these interactions. I thank Rajen Shah and Richard Samworth for their many insightful comments on my work when I interviewed with them. I look forward to our future collaborations in Cambridge. My thanks must go to my former mentor, Sangun Park, at Yonsei University who encouraged me to pursue a PhD degree. Without his advice and encouragement, I would not have made to this point.

I would also like to take this opportunity to thank many friends and colleagues who have made my five years in Pittsburgh wonderful. I first thank Kayla Frisoli and Xiao Hui Tai for their unconditional support and unwavering friendship over the last five years. We have so many wonderful memories together that I will never forget. I also thank my other cohort members: Neil Spencer, Jaehyeok Shin, Alden Green, Daren Wang, Yo Joong Choe for their support and help. Special thanks to my office mates, Neil Spencer and Alden Green, for the fruitful academic and non-academic conversations. I am also grateful to other

colleagues: Jisu Kim for grading my homework in great detail, Benjamin LeRoy for teaching me how to use the department servers, Robert Lunde for encouraging me to go to the gym, Niccolò Dalmasso for inviting me to the project for collaboration, Maria Jahja for buying me five points croissants, Kevin Lin for organizing our student seminar, Heejong Bong and Beomjo Park for having korean foods together, Yue Li and Wanshan Li for many late-night whiteboard discussions and Pratik Patil for numerous interesting questions. There are many other people including, but not limited to, Sangwon Hyun, Nick Kim, Kwangho Kim, Lee Richardson, Lingxue Zhu, Mikaela Meyer, Tudor Manole, Aleksandr Podkopaev, Jinjin Tian, Addison Hu, Tim Barry, Yufei Yi, Boyan Duan, Purvasha Chakravarti, Manjari Das, Xiaoyi Gu, Natalia Lombardi de Oliveira, Shamindra Shrotriya, Matteo Bonvini, Collin Politsch, Michael Stanley, Brendan McVeigh, Robin Dunn, Octavio Mesner, Shengming Luo, Minshi Peng, Zongge Liu, who have directly or indirectly contributed to this dissertation. I am so grateful to all of them.

Last but not least, I owe the most to my family, especially my parents to whom this thesis is dedicated, for their unconditional love and support.

To my parents

Abstract

With the recent advancement of data collection techniques, there has been an explosive growth in the size and complex of data sets in many application domains. The rise of such unprecedented data has posed new challenges as well as new opportunities to researchers in statistics and data science. Traditional methods, tailored to static and low-dimensional data, perform poorly or are no longer applicable for modern high-dimensional data with complex structures. Moreover, classical asymptotic theory easily breaks down under non-traditional settings where numerous parameters can interact in dynamic ways. Motivated by these new challenges, this dissertation aims to develop novel methods and technical tools suitable for modern high-dimensional data with particular emphasis on three types of testing problems: (i) one-sample testing, (ii) two-sample testing and (iii) independence testing.

One of the major contributions of this thesis is to introduce a flexible two-sample testing framework that can leverage any existing classification or regression method. By taking advantage of state-of-the-art algorithms in machine learning, the proposed method can efficiently handle different types of variables and various structures in high-dimensional data with competitive power under a variety of practical scenarios. To justify our approach, we provide rigorous theoretical and empirical analysis of their performance. With a specific focus on Fisher's linear discriminant analysis, we prove more sophisticated results including minimax optimality under common regularity conditions. In addition to supervised learning approaches, we also contribute to the literature by proposing goodness-of-fit tests for high-dimensional multinomials as well as multivariate generalizations of classical rank-based tests.

Another theme of this dissertation is concerned with permutation tests. Although the permutation approach is standard in practical implementations of two-sample and independence testing, its theoretical properties, especially power, have not been explored beyond simple cases. A major challenge of analyzing the permutation test is that it depends on a random critical value which is a function of observations. We study how to overcome this challenge and demonstrate that the permutation test has competitive power properties for many interesting problems under non-traditional settings. In particular we use the minimax perspective to evaluate the performance of a test and show that the permutation test is optimal for the problems where minimax lower bounds are available.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statements

Testing the equality of distributions is a fundamental topic in statistics with a wide range of applications. In astronomy, for example, researchers would like to explain underlying physical phenomenons based on their theoretical models. Testing whether the true distribution of physical objects agrees with a theoretical distribution helps astronomers decide the validity of their model approximation and obtain further insights into astronomical objects (Babu and Feigelson, 2006). In machine learning, it is of interest to determine whether the distribution of artificial images generated by an unsupervised learning method is similar to the underlying distribution of real images (Sutherland et al., 2016; Arjovsky et al., 2017). In marketing and business intelligence, the process of comparing two versions of a website or a mobile application, known as A/B testing, has been widely adopted to improve customer satisfaction and increase revenue (Siroker and Koomen, 2013).

Indeed, comparing distributions is a classical topic in statistics and there have been numerous methods developed since the pioneering work of Pearson (1900). Furthermore their theoretical and empirical properties have been well-established under classical low-dimensional regimes (e.g. Thas, 2010; Read and Cressie, 2012, for reviews). In recent years, however, we have witnessed renewed interest in this subject as the modern data we encounter are increasingly high-dimensional and complex (e.g. image data and network data). Traditional approaches — which focus on low-dimensional and Euclidean data — often fail or are not easily generalizable to high-dimensional and/or non-Euclidean data. For instance, classical Hotelling's T^2 test for the two-sample problem is only applicable when the dimension is less than the sample size and suffers from low power when the dimension and the sample size are comparable (Bai and Saranadasa, 1996). Some of the traditional approaches such as Kolmogorov–Smirnov test are based on empirical distributions and their extensions to multivariate cases are nontrivial. Motivated by these issues, the main goal of this thesis is to

propose statistical methods suitable for modern high-dimensional data and to understand their theoretical properties, particularly focusing on three hypothesis testing problems: 1) *the one-sample problem*, 2) *the two-sample problem* and 3) *the independence testing problem*. These problems can be formally stated as follows:

1. **One-sample problem.** Suppose we observe $\{X_1, \ldots, X_n\} \overset{i.i.d.}{\sim} P$ where $X_i \in \mathbb{R}^d$. Given a hypothesized distribution P_0, the one-sample problem aims at testing whether

$$H_0 : P = P_0 \quad \text{versus} \quad H_1 : P \neq P_0.$$

2. **Two-sample problem.** Let $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} P$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} Q$ where P and Q are unknown distributions and $X_i, Y_i \in \mathbb{R}^d$. The two-sample problem aims at testing whether

$$H_0 : P = Q \quad \text{versus} \quad H_1 : P \neq Q.$$

3. **Independence testing problem.** Given $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{i.i.d.}{\sim} P_{XY}$ where $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}^q$, we would like to test for independence between X and Y, i.e.

$$H_0 : P_{XY} = P_X P_Y \quad \text{versus} \quad H_1 : P_{XY} \neq P_X P_Y.$$

1.2 Permutation Approach

In testing problems, it is usual practice to determine a critical value in a way that the resulting test controls the type I error rate at level α. For the one-sample problem, finding such a critical value is not an issue as we can easily simulate the null distribution of a test statistic from known P_0. For the two-sample problem and the independence testing problem, however, the exact null distribution of the test statistic is rarely available as the underlying distributions are unknown. One common way to approximate this null distribution is based on asymptotic theory which presents a sharp result in large sample scenarios. This asymptotic approach, however, lacks a finite sample guarantee on type I error control and often requires stringent model assumptions that are hardly verifiable in practice. In theoretical computer science, on the other hand, testing problems are usually tackled from a non-asymptotic view where the established results hold in finite sample sizes building on concentration bounds (e.g. Canonne, 2015, for a survey). However the existing tests in this line of work are often impractical as their thresholds depend heavily on unspecified constants or even unknown parameters (e.g. Chan et al., 2014; Bhattacharya and Valiant, 2015; Diakonikolas and Kane, 2016). To bypass these issues, this thesis mainly adopts the permutation approach that yields a valid level α test for two-sample and independence testing (e.g. Pesarin and Salmaso, 2010; Good, 2013,

for reviews). While the permutation approach is standard in practical implementations of two-sample and independence testing, its theoretical properties, especially power, have not been explored beyond simple cases. A major challenge of analyzing the permutation test is that it depends on a random critical value which is a function of observations. One of the specific aims of this thesis is then to study how to overcome the issue arising from the random critical value and demonstrate that the permutation test has good power properties for many interesting problems under non-traditional settings.

1.3 Overview of this thesis

The rest of this thesis is organized as follows.

- **Chapter 2. Multinomial Goodness-of-Fit Based on U-Statistics:** In this chapter, we consider multinomial goodness-of-fit tests in the high-dimensional regime where the number of bins increases with the sample size. In this regime, Pearson's chi-squared test can suffer from low power due to the substantial bias as well as high variance of its statistic. To resolve these issues, we introduce a family of U-statistics for multinomial goodness-of-fit and study their asymptotic behaviors in high-dimensions. Specifically, we establish conditions under which the considered U-statistic is asymptotically Poisson or Gaussian, and investigate its power function under each asymptotic regime. Furthermore, we introduce a class of weights for the U-statistic that results in minimax rate optimal tests.

- **Chapter 3. Global and Local Two-Sample Tests via Regression:** The goal of this chapter is to present a regression approach to comparing multivariate distributions of complex data. Depending on the chosen regression model, our framework can efficiently handle different types of variables and various structures in the data, with competitive power under many practical scenarios. Whereas previous work has been largely limited to global tests which conceal much of the local information, our approach naturally leads to a local two-sample testing framework in which we identify local differences between multivariate distributions with statistical confidence. We demonstrate the efficacy of our approach both theoretically and empirically, under some well-known parametric and nonparametric regression methods. Our proposed methods are applied to simulated data as well as a challenging astronomy data set to assess their practical usefulness.

- **Chapter 4. Robust Multivariate Nonparametric Tests via Projection-Averaging:** In this chapter, we generalize the Cramér–von Mises statistic via projection-averaging to obtain a robust test for the multivariate two-sample problem. The proposed test is consistent against all fixed alternatives, robust to heavy-tailed data and minimax rate optimal against a certain class of alternatives. Our test statistic is completely free of tuning parameters and is computationally efficient even in high dimensions. When the dimension tends to infinity, the proposed test is shown to have comparable

power to the existing high-dimensional mean tests under certain location models. As a by-product of our approach, we introduce a new metric called *the angular distance* which can be thought of as a robust alternative to the Euclidean distance. Using the angular distance, we connect the proposed method to the reproducing kernel Hilbert space approach. In addition to the Cramér–von Mises statistic, we demonstrate that the projection-averaging technique can be used to define robust multivariate tests in many other problems.

- **Chapter 5. Comparing a Large Number of Multivariate Distributions:** In this chapter, we propose a test for the equality of multiple distributions based on kernel mean embeddings. Our framework provides a flexible way to handle multivariate data by virtue of kernel methods and allows the number of distributions to increase with the sample size. This is in contrast to previous studies that have been mostly restricted to classical univariate settings with a fixed number of distributions. By building on Cramér-type moderate deviation for degenerate two-sample V-statistics, we derive the limiting null distribution of the test statistic and show that it converges to a Gumbel distribution. The limiting distribution, however, depends on an infinite number of nuisance parameters, which makes it infeasible for use in practice. To address this issue, the proposed test is implemented via the permutation procedure and is shown to be minimax rate optimal against sparse alternatives.

- **Chapter 6. Euclidean and Manhattan Distance for High-Dimensional Two-Sample Testing:** The Euclidean distance is the most commonly used metric for high-dimensional data in the nonparametric two-sample testing literature. Many testing procedures based on Euclidean distance are known to be consistent for general alternatives under the classical low-dimensional setting. However, consistency becomes nontrivial for high-dimensional cases. In the high-dimension and low sample size setting, we demonstrate that existing nonparametric tests based on Euclidean distance become parametric tests. Specifically, we show under certain scenarios that they can be consistent only against first or second moment differences. We partially address this problem by replacing Euclidean distance with Manhattan distance.

- **Chapter 7. Classification accuracy as a proxy for two-sample testing:** When data analysts train a classifier and check if its accuracy is significantly different from chance, they are implicitly performing a two-sample test. We investigate the statistical properties of this flexible approach in the high-dimensional setting. We first present general conditions under which a classifier-based test is consistent, meaning that its power converges to one. To get a finer understanding of the rates of consistency, we study a specialized setting of distinguishing two Gaussians with different means and a common covariance. By focusing on Fisher's linear discriminant analysis (LDA) and its high-dimensional variants, we provide asymptotic but explicit power expressions of classifier-based tests and contrast them with corresponding Hotelling-type tests. Surprisingly, the expressions for their power

match exactly in terms of the parameters of interest, and the LDA approach is only worse by a constant factor.

- **Chapter 8. Minimax optimality of permutation tests:** In this chapter, we present a general non-asymptotic framework for analyzing the power of the permutation test. The utility of the proposed framework is illustrated in the context of two-sample and independence testing under both discrete and continuous settings. In each setting, we introduce permutation tests based on U-statistics and study their minimax performance. We also develop exponential concentration bounds for permuted U-statistics based on a novel coupling idea. Building on these exponential bounds, we introduce permutation tests which are adaptive to unknown smoothness parameters without losing much power. The proposed framework is further illustrated using more sophisticated test statistics including weighted U-statistics for multinomial testing and Gaussian kernel-based statistics for density testing.

Finally we conclude with a discussion of open problems and future directions in Chapter 9. Most of technical details and additional results are deferred to the appendices.

Chapter 2

Multinomial Goodness-of-Fit Based on U-Statistics: High-Dimensional Asymptotic and Minimax Optimality

This chapter is adapted from my work supervised by Sivaraman Balakrishnan and Larry Wasserman. This work was published in *Journal of Statistical Planning and Inference* (Kim, 2020).

2.1 Introduction

Suppose that there are n independent random vectors $\mathbf{X}_1 = (X_{1,1}, \ldots, X_{1,d}), \ldots, \mathbf{X}_n = (X_{n,1}, \ldots, X_{n,d})$ from a multinomial distribution with unknown parameters $\pi = (\pi_1, \ldots, \pi_d)$
$\in \Omega$ and

$$\Omega = \Big\{(\pi_1, \ldots, \pi_d) \in [0,1]^d : \sum_{j=1}^{d} \pi_j = 1\Big\}.$$

Given a specific choice of parameter vector $\pi_0 = (\pi_{0,1}, \ldots, \pi_{0,d}) \in \Omega$, the goodness-of-fit test for multinomial distributions is concerned with distinguishing

$$H_0 : \pi_0 = \pi \quad \text{versus} \quad H_1 : \pi_0 \neq \pi. \tag{2.1}$$

Pearson's chi-squared statistic (Pearson, 1900) is one of the well-known test statistics for this problem. Let $Y_j = \sum_{i=1}^n I(X_{i,j} = 1)$ for $j = 1, \ldots, d$. Then Pearson's chi-squared statistic is defined by

$$\chi_n^2 = \sum_{j=1}^d \frac{(Y_j - n\pi_{0,j})^2}{n\pi_{0,j}}.$$

The properties of χ_n^2 have been well-studied in a classical low-dimensional setting (Lehmann and Romano, 2006; Read and Cressie, 2012; Balakrishnan et al., 2013). For instance, the test based on χ_n^2 is asymptotically optimal against local alternatives when d is fixed (see, e.g. Chapter 14 of Lehmann and Romano, 2006). However, in the high-dimensional regime where the dimension is comparable with or much larger than the sample size, χ_n^2 suffers from the fact that it can have substantial bias for the testing problem. In other words, the power of the test can be much smaller than the significance level α against certain local alternatives. The major cause of the testing bias is due to the expected value of χ_n^2:

$$\mathbb{E}\left[\chi_n^2\right] = d - 1 + \sum_{j=1}^d \frac{\pi_j - \pi_{0,j}}{\pi_{0,j}} + \frac{n-1}{n} \sum_{j=1}^d \frac{(\pi_j - \pi_{0,j})^2}{\pi_{0,j}}.$$

When the null is not uniform, it is possible to observe $\mathbb{E}_{H_1}[\chi_n^2] < \mathbb{E}_{H_0}[\chi_n^2]$, which can trigger a significant bias problem of χ_n^2 for some α level. This bias problem becomes more serious when the dimension is large but the sample size is small (see, Haberman, 1988, for details).

To avoid the testing bias caused by the expected value, we view Pearson's chi-squared statistic as a V-statistic (Lemma 2.0.1) and consider a modified χ_n^2 based on the U-statistic. From the basic property of U-statistics, the modified χ_n^2 is an unbiased estimator of $\sum_{j=1}^d (\pi_j - \pi_{0,j})^2/\pi_{0,j}$ and its expectation becomes zero if and only if the null is true. As a result, the modified χ_n^2 can have significant power in the high-dimensional regime where classical χ_n^2 is substantially biased (Figure 2.1).

Another limitation of χ_n^2 in sparse multinomial settings is that it puts too much weight on small entries in π_0, and these small entries make the statistic highly unstable (Marriott et al., 2015; Valiant and Valiant, 2017; Balakrishnan and Wasserman, 2019). In this case, one might need to consider different weights to obtain higher power of the test. Motivated by these observations, we consider a family of U-statistics of $||A^{1/2}(\pi - \pi_0)||_2^2$ where A is some positive definite matrix.

The primary objective of this work is to investigate the limiting behavior of the proposed U-statistic in high-dimensions and determine a sufficient condition for A under which the resulting test is minimax rate optimal for multinomial goodness-of-fit.

Main results. The main results of this chapter are as follows:

1. *Poissonian Asymptotic for the U-statistic (Section 2.3.1)*: We establish conditions under which the U-statistic has a Poisson limiting distribution.

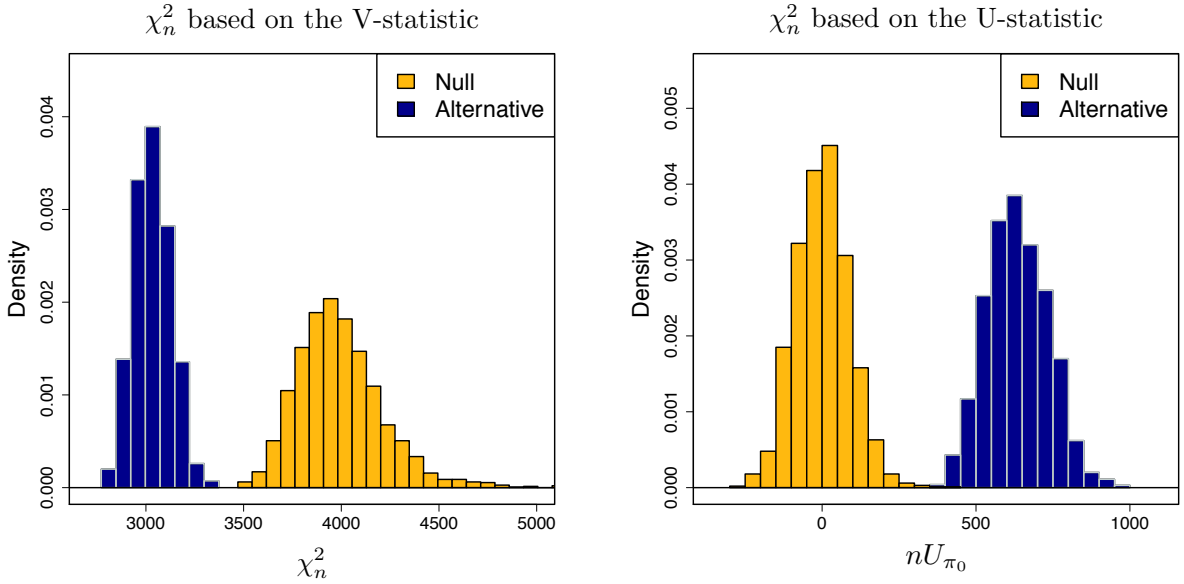

Figure 2.1: Illustration of the bias issue of Pearson's χ^2_n test in the high-dimensional regime. For the simulation, the null and the alternative are chosen $\pi_{0,i} \propto i$ and $\pi_i \propto i^5$, respectively. We take the sample size $n = 800$ and the dimension $d = 4000$. Since the null is rejected when the test statistic is greater than a certain quantile of the null distribution, we see from the left panel that the χ^2_n test is substantially biased. On the other hand, the right panel shows that the modified χ^2_n based on the U-statistic can have significant power in this example.

2. *Gaussian Asymptotic for the U-statistic (Section 2.3.2)*: We also establish conditions under which the U-statistic has a Gaussian limiting distribution.

3. *Global Minimax Optimality of the U-statistic (Section 2.4)*: We present a class of weight matrices A resulting in the minimax optimal test based on the U-statistic.

Related work. A considerable amount of literature has been published on the high-dimensional behavior of χ^2_n (e.g. Tumanyan, 1954, 1956; Steck, 1957; Holst, 1972; Morris, 1975; Read and Cressie, 2012; Rempała and Wesołowski, 2016, and the references therein). Our work is especially motivated by Rempała and Wesołowski (2016) who present conditions of the Poissonian and Gaussian asymptotics for χ^2_n. One can generalize their result to our U-statistic framework by using the relationship between U- and V-statistics. However, their analysis is restricted to the case of the null hypothesis and does not easily generalize to other cases with different weights. Hence, we take different approaches to overcome such shortcomings. The present study is also closely related to the work by Zelterman (1986, 1987) who proposes a modified χ^2_n to handle the testing

bias of the chi-squared test. The modified statistic is given by

$$\phi_n = \chi_n^2 - \sum_{j=1}^{d} \frac{Y_j}{n\pi_{0,j}}. \tag{2.2}$$

It can be shown that ϕ_n is equivalent to the proposed U-statistic up to some constant factors when we take the weight matrix as $A = \text{diag}(\pi_{0,1}^{-1}, \ldots, \pi_{0,d}^{-1})$ (Remark 2.1), and thus ϕ_n falls into our U-statistic framework. Diakonikolas et al. (2016) show that the collision-based test is minimax rate optimal for multinomial uniformity testing where $\pi_0 = (1/d, \ldots, 1/d)^\top$ (Remark 2.3). Their test statistic is a special case of our U-statistics by taking the identity weight matrix given in (2.6). We generalize their minimax result to an arbitrary null probability by providing a class of A that leads to the minimax optimal test. Our result includes the truncated weight considered in Balakrishnan and Wasserman (2019) as an example.

Outline. The rest of the paper is organized as follows. In Section 2.2, we view Pearson's chi-squared statistic as a V-statistic and provide a modified and generalized χ_n^2 based on the U-statistic. In Section 2.3, we investigate the Poissonian and Gaussian asymptotics for the proposed U-statistic in the high-dimensional regime. In Section 2.4, we present a sufficient condition for A that results in the minimax optimal test based on the U-statistic. In Section 2.5, we provide simulation studies. We summarize the results and discuss future work in Section 2.6. Finally, all of the proofs and additional results are presented in Appendix A.1.

2.2 Pearson's Chi-squared Statistic based on the U-statistic

As mentioned earlier, when the null is not uniform, Pearson's chi-squared statistic can have $\mathbb{E}_{H_1}[\chi_n^2] < \mathbb{E}_{H_0}[\chi_n^2]$. This phenomenon can result in serious testing bias especially in the high-dimensional regime. To remove the main testing bias due to its expected value, we first view χ_n^2 as a V-statistic and suggest the modified χ_n^2 based on a U-statistic. To begin, consider the following second order kernel:

$$h_{\pi_0}(\mathbf{X}_i, \mathbf{X}_j) = (\mathbf{X}_i - \pi_0)^\top D_{\pi_0}^{-1} (\mathbf{X}_j - \pi_0), \tag{2.3}$$

where D_{π_0} is a $d \times d$ diagonal matrix with the diagonal entries $\{\pi_{0,1}, \ldots, \pi_{0,d}\}$. Given h_{π_0}, we define the V-statistic as

$$V_{\pi_0} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_{\pi_0}(\mathbf{X}_i, \mathbf{X}_j).$$

In this setting, the next lemma shows that Pearson's chi-squared statistic is equivalent to the V-statistic defined with kernel h_{π_0}.

Lemma 2.0.1. *Pearson's chi-squared statistic has another representation as*

$$\chi_n^2 = \sum_{j=1}^{d} \frac{(Y_j - n\pi_{0,j})^2}{n\pi_{0,j}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} h_{\pi_0}(\mathbf{X}_i, \mathbf{X}_j) = nV_{\pi_0}.$$

As is well-known, a V-statistic is typically biased for estimating the population quantity. In order to remove the estimation bias of V_{π_0}, we consider a U-statistic defined as

$$U_{\pi_0} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h_{\pi_0}(\mathbf{X}_i, \mathbf{X}_j). \tag{2.4}$$

It can be easily seen that the expected value of U_{π_0} is always non-negative and equal to zero if and only if $\pi = \pi_0$. As a result, U_{π_0} does not suffer from the testing bias arising from the expected value.

Remark 2.1. *U_{π_0} is closely related to the test statistic proposed by Zelterman (1986, 1987). In view of Lemma 2.0.1, it is straightforward to show that these two statistics have the identity $\phi_n = (n-1)(U_{\pi_0} - 1)$; thus they have the exact same power. Unfortunately, even if U_{π_0} does not have the problem of the expectation, it can still have the testing bias against certain alternatives. For instance, Haberman (1988) provides a case where ϕ_n's power is less than the significance level, which implies that U_{π_0} also has the testing bias in the same case.*

There are some theoretical and empirical evidence to suggest that the scaling factor of χ_n^2 might not be optimal in the high-dimensional setting (Marriott et al., 2015; Valiant and Valiant, 2017; Balakrishnan and Wasserman, 2019). For example, when π_0 is highly skewed, χ_n^2 can perform poorly as it is dominated by small domain entries of π_0. Therefore, one might need to consider different weights for χ_n^2 to obtain higher power of the test in high-dimensions. In this context, we consider a family of U-statistics by considering a general weight matrix. The test statistic of our interest is given as

$$U_A = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h_A(\mathbf{X}_i, \mathbf{X}_j), \tag{2.5}$$

where $h_A(\mathbf{X}_i, \mathbf{X}_j) = (\mathbf{X}_i - \pi_0)^\top A(\mathbf{X}_j - \pi_0)$ and A is some positive definite matrix. In the next section, we study the limiting behavior of U_A under different high-dimensional regimes.

2.3 High-Dimensional Asymptotics

The asymptotic behavior of Pearson's chi-squared statistic has been well studied in the literature (Read and Cressie, 2012, for a review). In the high-dimensional case where the dimension (the number of bins) increases with the sample size, Rempała and Wesołowski (2016) investigate both the Poisson and Gaussian

approximations for χ^2_n. Specifically, when π_0 is uniformly distributed, they show that χ^2_n is asymptotically Poisson when $n/\sqrt{d} \to c \in (0, \infty)$ and asymptotically Gaussian when $n/\sqrt{d} \to \infty$. One can use their results to establish similar asymptotics for the U-statistic in view of Lemma 2.0.1. However, their analysis is restricted to the case of the null hypothesis.

In this section, we derive both null and alternative distributions of the considered U-statistic and present conditions for its high-dimensional limiting behavior. Note that, under the low-dimensional regime where d is fixed, it is rather straightforward to obtain the limiting distribution of the considered U-statistic. For example, U_A is the U-statistic, which has degeneracy of order one at the null hypothesis; thereby it converges to a weighted sum of independent chi-squared random variables with one degree of freedom (see, e.g., Lee, 1990, for asymptotic results of U-statistics). More interesting and challenging might be the high-dimensional case where U_A can have a Poisson or Gaussian limiting distribution, which will be studied in the following subsections.

2.3.1 Poisson Approximation

We start with establishing conditions under which the U-statistic has an asymptotic Poisson distribution. It is worth noting that, since a Poisson random variable is supported on the non-negative integers, an arbitrary choice of A does not necessarily yield a Poisson approximation for U_A (even after being properly centered and scaled). For this reason, we focus on the simple case where A is the identity matrix, i.e.

$$U_I = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} (\mathbf{X}_i - \pi_0)^\top (\mathbf{X}_j - \pi_0), \tag{2.6}$$

and study its asymptotic behavior. We briefly discuss the generalization of the identity matrix to an arbitrary matrix A resulting in the Poisson asymptotic in Remark 2.2.

Let us start by defining some conditions which hold as $n, d \to \infty$ simultaneously:

(P.1) $n^3 \{ \sum_{i=1}^d \pi_i^3 + (\pi^\top \pi)^2 \} \to 0$.

(P.2) $\binom{n}{2} \pi^\top \pi \to \eta_1$, $\binom{n}{2} \pi_0^\top \pi_0 \to \eta_0$ and $\binom{n}{2} \pi^\top \pi_0 \to \eta_2$ where $\eta_i \in (0, \infty)$ for $i = 0, 1, 2$.

(P.3) $n^3 \{ \sum_{i=1}^d \pi_i \pi_{0,i}^2 - \left(\sum_{i=1}^d \pi_i \pi_{0,i} \right)^2 \} \to 0$.

Let us consider the following decomposition of U_I:

$$\binom{n}{2} U_I = W - (n-1) \sum_{i=1}^n \left(\mathbf{X}_i^\top \pi_0 + \pi_0^\top \pi_0 \right),$$

where $W = \sum_{1 \le i < j \le n} \mathbf{X}_i^\top \mathbf{X}_j$. Based on this decomposition, first note that condition (P.1) together with the first condition in (P.2) is to ensure that W is approximately Poisson with mean η_1. The last two conditions

in (P.2) as well as (P.3) are to guarantee that the remainder of $\binom{n}{2}U_I$ other than W converges to $\eta_0 - 2\eta_2$ in probability. Specifically, (P.3) is a sufficient condition under which the variance of $(n-1)\sum_{i=1}^{n} \mathbf{X}_i^\top \pi_0$ converges to zero so that the asymptotic behavior of $\binom{n}{2}U_I$ is dominated by W.

Under the above conditions, we depict the limiting behavior of U_I as follows:

Theorem 2.1 (Poisson limiting distributions)**.** *Under the conditions* (P.1)*,* (P.2) *and* (P.3)*, U_I has a Poisson limiting distribution as*

$$\binom{n}{2}U_I \xrightarrow{d} \mathrm{Pois}(\eta_1) - 2\eta_2 + \eta_0.$$

In the following corollaries, we demonstrate the above result under the uniform null and the piecewise uniform alternatives.

Corollary 2.1.1 (Uniform null distribution)**.** *Suppose that we are under the uniform null, i.e. $\pi = \pi_0 = (1/d, \ldots, 1/d)^\top$. If $n/\sqrt{d} \to \sqrt{2\eta_0} \in (0, \infty)$, then*

$$\binom{n}{2}U_I \xrightarrow{d} \mathrm{Pois}(\eta_0) - \eta_0.$$

If $\eta_0 = 0$, then it converges to zero in distribution.

Corollary 2.1.2 (Piecewise uniform alternatives)**.** *Suppose that the null distribution is uniformly distributed. Consider $\omega_1, \omega_2 > 0$ such that $\omega_1 + \omega_2 = 1$. For simplicity, assume that d is even number (otherwise, let $\pi_{1,d} = 0$) and consider the alternative distribution defined by*

$$\pi_1 = (\underbrace{\omega_1/d, \ldots, \omega_1/d}_{d/2 \text{ elements}}, \underbrace{\omega_2/d, \ldots, \omega_2/d}_{d/2 \text{ elements}}).$$

If $n/\sqrt{d} \to \sqrt{2\eta_0} \in (0, \infty)$, then under the alternative hypothesis,

$$\binom{n}{2}U_I \xrightarrow{d} \mathrm{Pois}(\eta_1) - \eta_0,$$

where $\eta_1 = \eta_0(\omega_1^2 + \omega_2^2)/2$.

From the above results, let us describe the asymptotic power function of U_I under the Poissonian asymptotic. We assume that the null distribution is uniform where $n/\sqrt{d} \to c \in (0, \infty)$; thereby the distribution of $\binom{n}{2}(U_I + 1/d)$ is approximated by the Poisson distribution. Let $c_\alpha \in \mathbb{Z}^+$ be a critical value

such that

$$\mathbb{P}_{H_0}\left(\binom{n}{2}(U_I + 1/d) > c_\alpha\right) \leq \alpha,$$

under the null. Then the power function of U_I can be approximated by

$$\beta_{n,d}(\pi) = \mathbb{P}_{H_1}\left(\binom{n}{2}(U_I + 1/d) > c_\alpha\right) = \int_0^{2\eta_1} \frac{1}{\Gamma(c_\alpha + 1)} y^{c_\alpha} \exp\left(-\frac{y}{2}\right) dy + o(1), \tag{2.7}$$

against the alternatives that satisfy (P.1), (P.2) and (P.3).

Remark 2.2. *The Poisson approximation for U_I can be extended to a general statistic U_A when the weight matrix A is asymptotically close to σI for some $\sigma > 0$. Suppose that we are under the null hypothesis and the conditions (P.1), (P.2) and (P.3) are satisfied. For $\Sigma_0 = \mathrm{diag}(\pi_0) - \pi_0\pi_0^\top$ and $D_\sigma = A - \sigma I$, assume that $n^2\mathrm{tr}\{(D_\sigma\Sigma)^2\} \to 0$ as $n, d \to \infty$. Then the following holds by Chebyshev's inequality:*

$$\binom{n}{2}(U_A - \sigma U_I) \xrightarrow{p} 0 \quad \textit{and} \quad \binom{n}{2} U_A \xrightarrow{d} \sigma\mathrm{Pois}(\eta_0) - \sigma\eta_0.$$

2.3.2 Gaussian Approximation

In this section, we study the asymptotic normality of U_A. Without loss of generality, we further assume that A is symmetric. Under the null hypothesis, the next theorem provides a sufficient condition under which U_A is asymptotically Gaussian.

Theorem 2.2 (Asymptotic normality of U_A under the null)**.** *Suppose*

$$\frac{\mathrm{tr}((A\Sigma)^4)}{[\mathrm{tr}\{(A\Sigma)^2\}]^2} \to 0 \quad \textit{and} \quad \frac{\mathbb{E}\big[\{h_A(\boldsymbol{X}_1, \boldsymbol{X}_2)\}^4\big] + n\mathbb{E}\big[\{h_A(\boldsymbol{X}_1, \boldsymbol{X}_2)\}^2\{h_A(\boldsymbol{X}_1, \boldsymbol{X}_3)\}^2\big]}{n^2[\mathrm{tr}\{(A\Sigma)^2\}]^2} \to 0. \tag{2.8}$$

Then, under the null hypothesis, we have

$$\sqrt{\binom{n}{2}} \frac{U_A}{\sqrt{\mathrm{tr}\big\{(A\Sigma)^2\big\}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Recall that, in Corollary 2.1.1, we established the limiting behavior of U_I under the uniform null when $n/\sqrt{d} \to c \in [0, \infty)$. In the next corollary, we study the asymptotic normality of U_I when $n/\sqrt{d} \to \infty$ under the uniform null.

Corollary 2.2.1 (Uniform null distribution). *Suppose we are under the null hypothesis where π_0 is uniform. If $n/\sqrt{d} \to \infty$, then we have*

$$\sqrt{\binom{n}{2}} \frac{U_I}{\sqrt{1/d(1-1/d)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Let us now turn our attention to the alternative distribution of U_A in the high-dimensional asymptotic. As in Chen and Qin (2010), we consider the following two scenarios under the alternative hypothesis:

(S.1) (Strong Signal-to-Noise) $n^{-1}\mathrm{tr}((A\Sigma)^2) = o\big((\pi-\pi_0)^\top A\Sigma A(\pi-\pi_0)\big)$.

(S.2) (Weak Signal-to-Noise) $(\pi-\pi_0)^\top A\Sigma A(\pi-\pi_0) = o\big(n^{-1}\mathrm{tr}((A\Sigma)^2)\big)$.

To appreciate the given scenarios, let us decompose $U_A = U_{\text{quad}} + U_{\text{linear}}$ where

$$U_{\text{quad}} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} (\mathbf{X}_i - \pi)^\top A(\mathbf{X}_j - \pi),$$

$$U_{\text{linear}} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \Big\{(\mathbf{X}_i - \pi)^\top A(\pi-\pi_0) + (\mathbf{X}_j - \pi)^\top A(\pi-\pi_0) + ||A^{1/2}(\pi-\pi_0)||_2^2\Big\}.$$

Accordingly, we have $\mathbb{E}[U_{\text{quad}}] = 0$ and $\mathbb{E}[U_{\text{linear}}] = ||A^{1/2}(\pi-\pi_0)||_2^2$. Under (S.1), U_A is dominated by U_{linear} and thus

$$\frac{U_A - ||A^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{\mathrm{Var}(U_A)}} = \frac{U_{\text{linear}} - ||A^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{\mathrm{Var}(U_{\text{linear}})}} + o_P(1),$$

whereas under (S.2), U_A is dominated by U_{quad} so that

$$\frac{U_A - ||A^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{\mathrm{Var}(U_A)}} = \frac{U_{\text{quad}}}{\sqrt{\mathrm{Var}(U_{\text{quad}})}} + o_P(1).$$

Hence, in order to establish the asymptotic normality of U_A, we need to study the limiting behavior of U_{linear} and U_{quad} under each scenario. The result is summarized in the following theorem.

Theorem 2.3 (Asymptotic normality of U_A under the alternative). *Assume either i)* (S.1) *and* $(\pi-\pi_0)^\top A\Sigma A(\pi-\pi_0) < \infty$, *or ii)* (S.2) *and the condition (2.8) given in Theorem 2.2. Then*

$$\frac{U_A - ||A^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{\mathrm{Var}(U_A)}} \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\mathrm{Var}\,(U_A) = \binom{n}{2}^{-1}\Big\{\mathrm{tr}\{(A\Sigma)^2\} + 2(n-1)(\pi-\pi_0)^\top A\Sigma A(\pi-\pi_0)\Big\}.$$

Theorem 2.3 together with Theorem 2.2 allows us to describe the power function of U_A under the Gaussian asymptotic. Let z_α be the upper α quantile of the standard normal distribution. For notational simplicity, let us denote

$$\Lambda_0 = \mathrm{tr}\{(A\Sigma_0)^2\}, \quad \Lambda_1 = \mathrm{tr}\{(A\Sigma)^2\} \quad \text{and} \quad \Lambda_2 = 2(n-1)(\pi-\pi_0)^\top A\Sigma A(\pi-\pi_0),$$

where Σ_0 and Σ are the covariance matrix of $\mathbf{X}$ under the null and the alternative hypothesis, respectively. Then the power is approximated by

$$\beta_{n,d}(\pi_0,\pi_1,A) = \Phi\left(-\frac{\sqrt{\Lambda_0}}{\sqrt{\Lambda_1+\Lambda_2}}z_\alpha + \sqrt{\binom{n}{2}}\frac{||A^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{\Lambda_1+\Lambda_2}}\right) + o(1). \tag{2.9}$$

Under (S.1) together with the additional assumption (S.3) below:

(S.3) $n^{-1}\mathrm{tr}((A\Sigma_0)^2) = o\big((\pi-\pi_0)^\top A\Sigma A(\pi-\pi_0)\big)$,

the power function of U_A can be further simplified to

$$\beta_{n,d}(\pi_0,\pi_1,A) = \Phi\left(\frac{\sqrt{n}||A^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{4(\pi-\pi_0)^\top A\Sigma A(\pi-\pi_0)}}\right) + o(1).$$

On the other hand, under (S.2), the approximation becomes

$$\beta_{n,d}(\pi_0,\pi_1,A) = \Phi\left(-\frac{\sqrt{\mathrm{tr}\{(A\Sigma_0)^2\}}}{\sqrt{\mathrm{tr}\{(A\Sigma)^2\}}}z_\alpha + \frac{n||A^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{2\mathrm{tr}\{(A\Sigma)^2\}}}\right) + o(1).$$

2.4 Minimax Optimality

As discussed before, χ^2 statistic tends to have a large variance by putting too much weight on small entries of π_0. Consequently, the resulting test can perform poorly and is not minimax optimal in the high-dimensional setting (Balakrishnan and Wasserman, 2018). This motivates us to consider different weights for the test statistic. In this section, we discuss the choice of the weight matrix A from a minimax point of view. To formulate the minimax problem, we modify the hypotheses given in (2.1) as

$$H_0 : \pi = \pi_0 \quad \text{versus} \quad H_1 : ||\pi-\pi_0||_1 \geq \epsilon_n, \tag{2.10}$$

where $||x||_1$ is the ℓ_1 norm of $x \in \mathbb{R}^d$. Let us consider a set of level α test functions, $\phi : \{\mathbf{X}_i\}_{i=1}^n \mapsto \{0,1\}$, such that

$$\Phi_{n,\alpha} = \Big\{\phi : \mathbb{P}_{H_0}^n (\phi = 1) \leq \alpha, 0 < \alpha < 1\Big\}. \tag{2.11}$$

Then the global minimax risk (see e.g., Valiant and Valiant, 2017; Balakrishnan and Wasserman, 2019) is defined as the supremum over the local minimax risk:

$$R_n(\epsilon_n) = \sup_{\pi_0 \in \Omega} R_n(\epsilon_n, \pi_0),$$

where the local minimax risk is given by

$$R_n(\epsilon_n, \pi_0) = \inf_{\phi \in \Phi_{n,\alpha}} \sup \Big\{\mathbb{E}_{H_1}[1 - \phi] : ||\pi - \pi_0||_1 \geq \epsilon_n, \pi \in \Omega\Big\}.$$

For a given $\delta \in (0, 1 - \alpha)$, the global minimum separation rate is characterized by

$$\epsilon_n^* = \inf \Big\{\epsilon_n : R_n(\epsilon_n) \leq \delta\Big\}.$$

Under the given setting, Valiant and Valiant (2017) show that the global minimax rate is

$$\epsilon_n^* \asymp \frac{d^{1/4}}{\sqrt{n}}.$$

The main objective of this section is to find a sufficient condition for A that results in minimax rate optimal test based on the U-statistic. We first describe that the test based on the U-statistic with a mixture weight is minimax rate optimal in Section 2.4.1 and we go on to generalize this result in Section 2.4.2.

2.4.1 U-statistic weighted by a mixture distribution

The weight used in U_{π_0} often results in a high variance of the test statistic especially when π_0 is sparse. U_I does not suffer from the same variance issue but its weight does not use information of the null distribution. We combine U_{π_0} and U_I to reduce the disadvantages associated with each and obtain a minimax rate optimal test. Let us define the mixture distribution by

$$\pi_{\text{mix}} = \frac{1}{2}\pi_0 + \frac{1}{2}\pi_{\text{unif}}, \tag{2.12}$$

where $\pi_{\text{unif}} = (1/d, \ldots, 1/d)$. Then by using $A_{\text{mix}} = \text{diag}\{\pi_{\text{mix},1}^{-1}, \ldots, \pi_{\text{mix},d}^{-1}\}$, the resulting U-statistic is defined by

$$U_{\text{mix}} = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} (\mathbf{X}_i - \pi_0)^\top A_{\text{mix}} (\mathbf{X}_j - \pi_0). \tag{2.13}$$

To test (2.10), we reject the null hypothesis when U_{mix} is greater than the critical value:

$$\phi(U_{\text{mix}}) \;=\; I\left(U_{\text{mix}} > \sqrt{\frac{1}{\alpha}\binom{n}{2}^{-1} \text{tr}\{(A_{\text{mix}}\Sigma_0)^2\}} \right),$$

where $\Sigma_0 = \text{diag}\,(\pi_0) - \pi_0\pi_0^\top$. Then we can see that $\phi(U_{\text{mix}})$ has nontrivial power when $\epsilon_n \asymp d^{1/4}/\sqrt{n}$ and thus it is global minimax rate optimal. We formally state this result in Theorem 2.4 which holds for more general test statistics.

Remark 2.3. *Diakonikolas et al. (2016) show that the collision-based test statistic*

$$W = \sum_{1 \le i < j \le n} \mathbf{X}_i^\top \mathbf{X}_j$$

is minimax rate optimal for multinomial uniformity testing. For the uniform null case, U_{mix} is equivalent to the collision-based test statistic W; thereby, our result can be viewed as a generalization of Diakonikolas et al. (2016) to arbitrary null probabilities.

Remark 2.4. *Poissonization, where the sample size has a Poisson distribution, is a standard assumption in the literature to construct the upper bound of the minimax risk. Under Poissonization, several statistics have been proposed to obtain the minimax optimality (Valiant and Valiant, 2017; Balakrishnan and Wasserman, 2019). We would like to emphasize that our minimax result is established without assuming Poissonization.*

2.4.2 Generalization

The mixture distribution in (2.12) can be generalized by considering an arbitrary but fixed $\gamma \in (0, 1)$ such that

$$\pi_{\text{mix}}^{(\gamma)} = \gamma\pi_0 + (1-\gamma)\pi_{\text{unif}}.$$

For a given weight vector $w \in \mathbb{R}^d$, we say that $w \in \mathbb{R}^d$ is *comparable* to $\pi_{\text{mix}}^{(\gamma)}$, if there exist fixed constants $C_1, C_2 > 0$ independent of n and d such that $C_1\pi_{\text{mix},i}^{(\gamma)} \le w_i \le C_2\pi_{\text{mix},i}^{(\gamma)}$ for all $i = 1, \ldots, d$. We denote a

weight vector comparable to $\pi_{\text{mix}}^{(\gamma)}$ by

$$w \sim \pi_{\text{mix}}^{(\gamma)}.$$

Based on these notations, let us define a class of weight matrices:

$$\mathcal{A}_w = \left\{ \text{diag}\left(w^{-1}\right) \in \mathbb{R}^{d\times d} : w \in \mathbb{R}^d \text{ and } w \sim \pi_{\text{mix}}^{(\gamma)} \text{ for some } \gamma \in (0,1) \right\}. \tag{2.14}$$

Then the test based on the U-statistic associated with any $A_w \in \mathcal{A}_w$:

$$\phi(U_w) \;=\; I\left(U_w > \sqrt{\frac{1}{\alpha}\binom{n}{2}^{-1} \text{tr}\{(A_w\Sigma_0)^2\}} \right)$$

is global minimax rate optimal. The result is summarized in the next theorem.

Theorem 2.4 (Global minimax optimality of U_w)**.** *For testing (2.10), the test based on $\phi(U_w)$ has size at most α. In addition, suppose there exists a universal constant $C > 0$ independent of n and d such that*

$$\epsilon_n^2 \geq \frac{C\sqrt{d}}{n}\left[\frac{1}{\sqrt{\alpha}} + \frac{1}{\zeta}\right], \tag{2.15}$$

for any $\zeta \in (0,1]$, then we have $\mathbb{P}_{H_1}(\phi(U_w) = 0) \leq \zeta$. Hence, $\phi(U_w)$ is global minimax optimal.

Here we provide several examples that belong to the proposed framework.

Example 2.1 (Truncated χ^2)**.** *Balakrishnan and Wasserman (2019) show that the test based on the truncated χ^2 test statistic is global minimax rate optimal. Unlike the classical χ^2 statistic, the truncated χ^2 test statistic is weighted by $\theta_{\text{trunc},j} = \max\{\pi_{0,j}, 1/d\}$ for $j = 1,\ldots,d$. Note that $\theta_{\text{trunc}} \sim \pi_{\text{mix}}$ since $\pi_{\text{mix},j} \leq \theta_{\text{trunc},j} \leq 2\pi_{\text{mix},j}$ for all $j = 1,\ldots,d$. Therefore, it satisfies the comparable condition with $C_1 = 1$ and $C_2 = 2$.*

Example 2.2 (ℓ_p-type mixture)**.** *For $p \geq 1$, let us define*

$$\theta_{\ell_p,j} = \left(\frac{\pi_{0,j}^p + \pi_{\text{unif},j}^p}{2}\right)^{1/p}$$

for $j = 1,\ldots,d$. Then we observe that $\theta_{\ell_p} \sim \pi_{\text{mix}}$ since $\pi_{\text{mix},j} \leq \theta_{\ell_p,j} \leq 2^{1-1/p}\pi_{\text{mix},j}$ for all $j = 1,\ldots,d$, where we used $||x||_1 \leq 2^{1-1/p}||x||_p \leq 2^{1-1/p}||x||_1$ for $p \geq 1$. In fact, if $p = \infty$, it corresponds to the truncated weight as $\theta_{\ell_\infty} = \theta_{\text{trunc}}$.

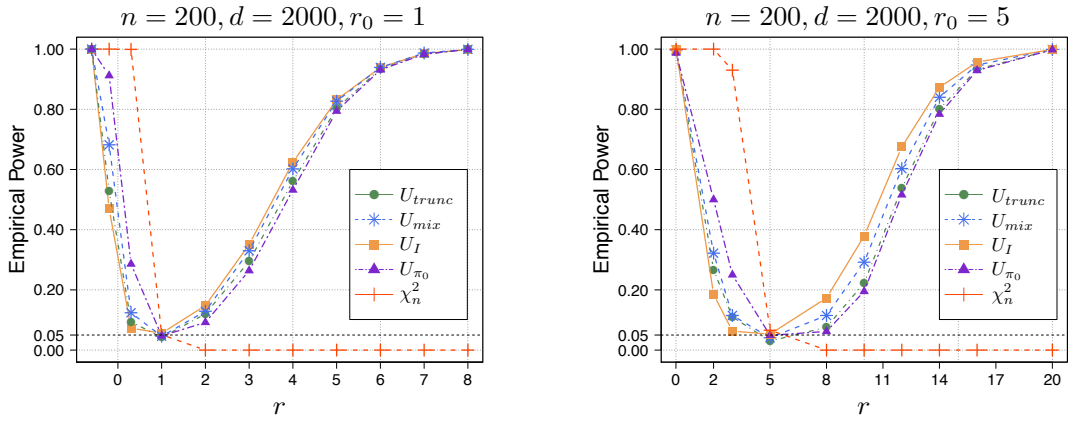

Figure 2.2: Power comparisons between five different tests based on U_{trunc}, U_{mix}, U_I, U_{π_0} and Pearson's χ^2_n at significance level $\alpha = 0.05$.

2.5 Simulations

In this section, we provide numerical results to illustrate the finite sample performance of the proposed methods. In the first simulation study, we compare power between Pearson's chi-squared test and the proposed tests based on U-statistics. We consider four different U-statistics: U_{π_0}, U_I, U_{mix} and U_{trunc} where U_{trunc} is the U-statistic with truncation weights described in Example 2.1. We let the null distribution π_0 have a power law distribution where the probability of the ith bin is proportional to the r_0th power of its index, i.e. $\pi_{0,i} \propto i^{r_0}$ for $i = 1, \ldots, d$. When r_0 is close to zero, then the null distribution becomes close to the uniform distribution. On the other hand, when r_0 has a large value, the null distribution becomes skewed to the left. In our simulations, we consider two null distributions with $r_0 = 1$ and $r_0 = 5$. The alternative distribution π is also chosen to have a power law distribution as $\pi_i \propto i^r$ for $i = 1, \ldots, d$ and we change r to describe different power behaviors. The null and alternative distribution of each statistic are estimated via Monte Carlo simulations with 1000 repetitions where we take the sample size and the number of bins as $n = 200$ and $d = 2000$, respectively.

The simulation results are presented in Figure 2.2. From the results, we observe that Pearson's χ^2_n test shows entirely different behaviors between two alternatives where (i) $r_0 < r$ and (ii) $r_0 > r$. Specifically, when $r_0 < r$, Pearson's χ^2_n test is extremely biased and has zero power to reject the null hypothesis, whereas it has the highest power among the considered tests when $r_0 > r$. In contrast, the tests based on the U-statistics are considerably robust toward the testing bias and perform reasonably well against the entire range of alternatives. This illustrates the benefit of the proposed U-statistic framework under the high-dimensional regime. For the comparison between the U-statistics, no test is uniformly more powerful than the others. In particular, the test based on U_{π_0} outperforms the other tests when $r_0 > r$, but underperforms when $r_0 < r$.

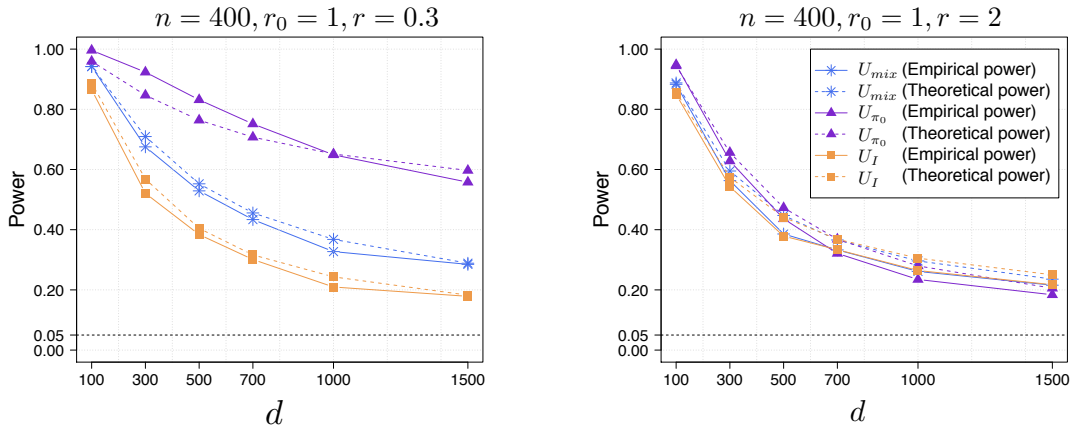

Figure 2.3: Comparisons between the empirical power and the theoretical power based on the normal approximation at significance level $\alpha = 0.05$.

On the other hand, the test based on U_I performs the best when $r_0 < r$ and performs the worst when $r_0 > r$. The test based on U_{mix} are usually the second best and better than the one based on U_{trunc}.

In the second simulation study, we compare the empirical power of the tests and the corresponding theoretical power based on the normal approximation established in (2.9). For the comparison, we consider three U-statistics: U_{π_0}, U_I and U_{mix}, and choose the null distribution as $\pi_{0,i} \propto i$ for $i = 1, \ldots, d$. Under the alternative, we consider two power law distributions with $r = 0.3$ and $r = 2$ where $\pi_i \propto i^r$ for $i = 1, \ldots, d$. As before, the null and alternative distribution of each statistic are estimated via Monte Carlo simulations with 1000 repetitions where we take the sample size and the number of bins as $n = 400$ and $d = 100, 300, 500, 700, 1000, 1500$. The results are given in Figure 2.3. It is seen from the results that the power approximation via asymptotic normality looks fairly robust over different dimensions especially against the alternative distribution with $r = 2$.

2.6 Summary and Discussion

In this work, we introduced a family of U-statistics for multinomial goodness-of-fit tests and investigated their asymptotic behaviors in the high-dimensional regime. Specifically, we established the conditions under which the U-statistic is approximately Poisson or Gaussian, and studied its power function under each asymptotic regime. We also proposed a class of weights for the U-statistic and showed the minimax optimality of the resulting tests. Despite the fact that the proposed tests achieve minimax rate optimality, they still have room for improvement. In particular, the considered class of weight functions only uses the information of π_0 but not π. When prior information about π is available (e.g. differences exist in specific bins with high probability), then it is possible to design more powerful test by incorporating that information. In this case,

it would be beneficial to use our asymptotic results and choose A that maximizes the asymptotic power function under the given restrictions. We reserve this topic for future work.

Chapter 3

Global and Local Two-Sample Tests via Regression

This chapter is adapted from my joint work with Ann Lee and Jing Lei. This work was published in *Electronic Journal of Statistics* (Kim et al., 2019a).

3.1 Introduction

Given two distributions P_0 and P_1 on $\mathbb{R}^D$, the global two-sample problem is concerned with testing $H_0 : P_0 = P_1$ versus $H_1 : P_0 \neq P_1$, based on independent random samples from each distribution. This fundamental problem has a long history in statistics and has been well-studied in a classical setting (see, e.g., Thas, 2010). Recently, however, there has been renewed interest in this field as modern data we encounter have become more complex and diverse. Traditional approaches, which focus on low-dimensional and Euclidean data, often fail or are not easily generalizable to high-dimensional and non-Euclidean data. Additionally, some recent developments in high-dimensional two-sample testing are limited to simple alternatives such as location and scale differences (see, Hu and Bai, 2016, for a recent review). In this context, there is a need to develop a new tool for the two-sample problem that can efficiently handle complex data and can detect differences beyond location and scale alternatives.

When the null hypothesis of the global two-sample test is rejected, it is often valuable (for e.g. scientific discovery, calibration of simulation models, and so on) to further explore *how* the two distributions are different. Specifically, as a follow-up study to the global test, one might wish to identify locally significant regions where the two distributions differ. This topic, which we refer to as the *local two-sample problem*, has been studied by Duong (2013) who uses kernel density estimators to identify local differences between two

density functions. However, the kernel density approach may perform poorly when distributions are not in a low-dimensional Euclidean space, and hence another tool is needed for more challenging settings.

The goal of this work is to develop a general framework for both global and local two-sample problems that overcomes the aforementioned challenges. Specifically, we aim to design a two-sample test that can efficiently handle different types of variables (e.g. mixed data types) and various structure (e.g. manifold, irrelevant covariates) in the data. Consequently, the resulting test can have substantial power for a variety of challenging alternatives. We achieve our goal by connecting the two-sample problem to a regression problem as follows. Let f_0 and f_1 be density functions of P_0 and P_1 with respect to a common dominating measure. We view f_0 and f_1 as conditional densities $f(x|Y = 0)$ and $f(x|Y = 1)$ by introducing an indicator random variable $Y \in \{0, 1\}$. Then by Bayes' theorem, the hypothesis $H_0 : f_0(x) = f_1(x)$ for all $x \in S = \{x \in \mathbb{R}^D : f_0(x) + f_1(x) > 0\}$ can be verified to be equivalent to the hypothesis that involves the regression function:

$$H_0 : \ \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1), \quad \text{for all } x \in S. \tag{3.1}$$

We state the corresponding global and local alternative hypotheses as

$$H_1 : \ \mathbb{P}(Y = 1|X = x) \neq \mathbb{P}(Y = 1), \quad \text{for some } x \in S, \text{ and}$$

$$H_1(x) : \ \mathbb{P}(Y = 1|X = x) \neq \mathbb{P}(Y = 1), \quad \text{at fixed } x \in S,$$

respectively.

Motivated by the above reformulation, we propose a testing procedure that measures an empirical distance between the regression function $\mathbb{P}(Y = 1|X = x)$ and the class probability $\mathbb{P}(Y = 1)$. We refer to this approach as *the regression test*. Depending on the choice of regression method, the regression test can adapt to nontraditional data settings. As we shall see, the power of the test is closely related to the mean square error of the chosen regression estimator. In addition, by choosing a nonparametric regression method, the global regression test can be sensitive to general alternatives beyond location and scale differences. We will demonstrate the benefits of the regression test with both theoretical and empirical results.

3.1.1 Motivating Example

We motivate our approach by comparing multivariate distributions of galaxy morphologies, but the proposed framework benefit other areas of science and technology as well (involving, e.g., outlier detection, calibration of simulation models, and comparison of cases and controls). A galaxy's morphology is the organization of a galaxy's light, as projected into our line of sight and observed at a particular wavelength as a pixelated image. Morphological studies are key to understanding the evolutionary history of galaxies and to constraining

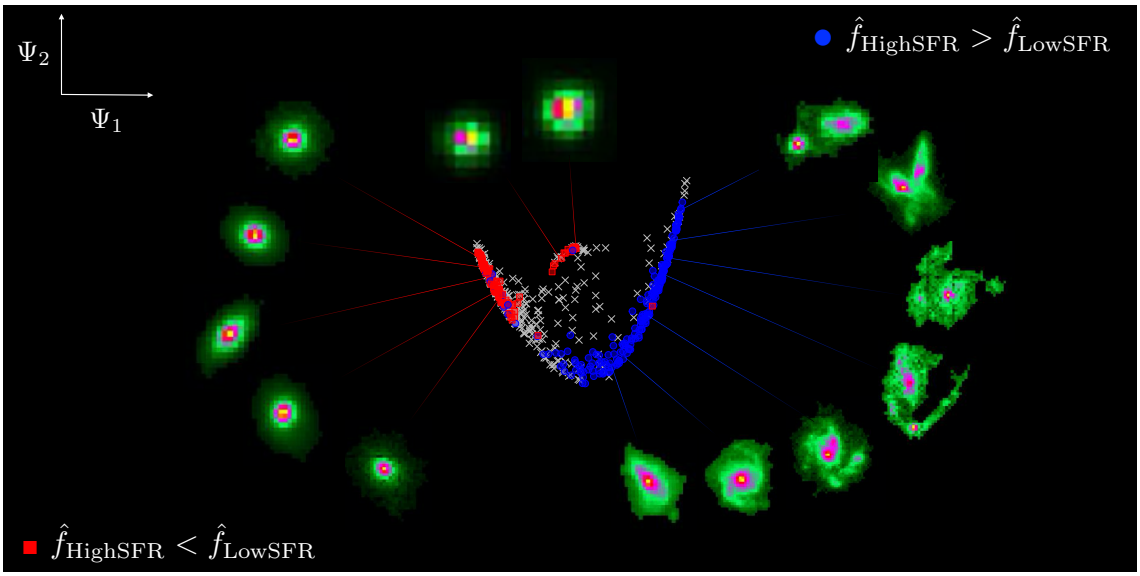

Figure 3.1: Result of local two-sample test of differences between high- and low-SFR galaxies in a seven-dimensional morphology space. The red squares indicate regions where the density of low-star-forming galaxies are significantly higher, and the blue circles indicate regions in morphology space that are dominated by high-star-forming galaxies; the gray crosses represent insignificant test points. The galaxies are embedded in a two-dimensional diffusion space for visualization purposes only (see Appendix B.2 for details); Ψ_1 and Ψ_2 here denote the first two coordinates.

theories of the Universe; see e.g. Conselice (2014) for a review. So far astronomers have only been able to study one or two morphological statistics (or projections of these) at a time instead of an entire ensemble. The reason is a lack of tools for effectively comparing and jointly analyzing multivariate or high-dimensional data in their native spaces. A global hypothesis test with a binary reject yes/no answer is also not informative enough to explain how two distributions are different in a multivariate feature space.

We illustrate the efficacy of the proposed global and local testing framework on the morphology statistics of two galaxy populations with high and low star-formation rate (SFR), respectively. The challenge here is not only that the problem involves multivariate data, but also that some of the morphological statistics are mixed discrete and continuous type with heavy outliers. We efficiently handle this issue by building on the success of random forests regression. The visualized local two-sample result is shown in Figure 3.1 and the details of the analysis can be found in Section 3.6.

3.1.2 Related Work

In recent years, several attempts have been made to connect binary classification with two-sample testing. The main idea of this approach is to check whether the accuracy of a binary classifier is better than chance level and reject the null if the difference is significant. Such an approach, referred to as an accuracy or

classification test, was conceptualized by Friedman (2003) and has since been investigated by several authors (Ojala and Garriga, 2010; Olivetti et al., 2015; Ramdas et al., 2016; Rosenblatt et al., 2016; Gagnon-Bartsch and Shem-Tov, 2019; Lopez-Paz and Oquab, 2016; Hediger et al., 2019). In the same manner as our regression framework, a key strength of the accuracy test is that it offers a flexible way for the two-sample problem as it can utilize any existing classification procedure in the literature. However, the classification accuracy framework is not easily converted to a local two-sample test. In addition, many classifiers are estimated by dichotomizing regression estimators and the discrete nature of such classifiers may result in a less powerful test (see Section 3.5.2 and other simulation results).

For the global two-sample test, our framework can be viewed as an instance of goodness-of-fit testing for regression models (e.g. González-Manteiga and Crujeiras, 2013, for a review). There is a substantive literature on this topic including Hardle and Mammen (1993), Weihrather (1993), González-Manteiga and Cao (1993), Zheng (1996), Zhang and Dette (2004), Hart (2013) and among others. This line of work typically concentrates on comparing differences between parametric (e.g. linear regression) and nonparametric (e.g. kernel regression) fits from an asymptotic point of view. For example, Hardle and Mammen (1993) consider the squared deviation between a parametric regression estimator and a kernel estimator. They show that their test statistic converges to a normal distribution under the null hypothesis and justify the use of the wild bootstrap procedure. However, this type of asymptotic approach is challenging to analyze beyond kernel-type estimators and often requires strong technical assumptions. In contrast, our framework is designed to compare any type of regression estimators with a specific constant fit by building upon the permutation principle. Hence the resulting test is valid in any finite sample sizes. Moreover we present a unified framework of studying the power of the regression test by taking advantage of existing results on the estimation error.

For the local two-sample test, our approach has similarities to independent work by Cazáis and Lhéritier (2015) who estimate the Kullback-Leibler divergence between $\mathbb{P}(Y = 1|X = x)$ and $\mathbb{P}(Y = 1)$. Our procedure, however, identifies locally significant areas with statistical confidence whereas Cazáis and Lhéritier (2015) graphically decide a threshold for the significance.

3.1.3 Overview of this chapter

We outline this chapter as follows: In Section 3.2, we introduce the proposed metrics, test statistics and algorithms for the global and local regression tests. In Section 3.3, we study theoretical properties of the global regression test. We begin by considering a simple scenario where two populations only differ in their means in Section 3.3.1. In this scenario, we show that the regression test based on Fisher's linear discriminant analysis (LDA) achieves the same local optimality as the Hotelling's T^2 test. Moving on to general regression settings in Section 3.3.2, we establish a connection between the testing error of the global regression test

and the mean integrated square error (MISE) of the regression estimator. In Section 3.4, we turn to the local two-sample problem and investigate general properties of the local regression tests. In Section 3.4.1, we describe the testing error of the local regression test in terms of the mean square error (MSE) of the regression estimator. We further establish an optimality of the local regression tests over the Lipschitz class from a minimax point of view in Section 3.4.2. When data have intrinsic dimension, we show that the performance of the local regression tests based on kNN or kernel regression only depends on intrinsic dimension in Section 3.4.3. Section 3.4.4 studies the limiting distribution of the local permutation statistic to avoid a high computational cost from permutations for large sample size. In Section 8.9, simulation studies are provided to illustrate finite sample performance of the global and local regression tests. In Section 3.6, we apply the proposed approach to a problem in astronomy and demonstrate its efficacy. All the proofs are deferred to Appendix C.4.

Notation. Throughout this chapter, we denote the class probabilities $\mathbb{P}(Y = 0)$ and $\mathbb{P}(Y = 1)$ by π_0 and π_1, respectively, and write the joint distribution of (X, Y) by $\pi_0[P_0 \times \delta_0] + \pi_1[P_1 \times \delta_1]$ where δ_k denotes the point mass at k for $k = 0, 1$. We denote the corresponding conditional probability $\mathbb{P}(Y = 1|X = x)$ by $m(x)$, which can be explicitly written as

$$m(x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)}.$$

We use $P_X(\cdot)$ to denote the marginal probability measure of X and $||Z||_2$ denotes the Euclidean norm of a vector $Z \in \mathbb{R}^D$. The symbols $\xrightarrow{p}$ and $\xrightarrow{d}$ stand for convergence in probability and in distribution, respectively. We use $a_n \lesssim b_n$ if there exists $C > 0$ such that $a_n \leq Cb_n$ for all n. Similarly, $a_n \asymp b_n$ if there exist constants $C, C' > 0$ such that $C \leq |a_n/b_n| \leq C'$ for all n. As convention, the acronym *i.i.d.* is used to represent independent and identically distributed.

3.2 Framework

3.2.1 Metrics

A common metric for comparing two distributions is the difference between two density functions $f_0(x)$ and $f_1(x)$; this metric has been used for global and local two-sample testing by Anderson et al. (1994) and Duong (2013). Another natural metric, suggested for global two-sample testing by Keziou and Leoni-Aubin (2005), Fokianos (2008) and Sugiyama et al. (2011), is the density ratio $f_1(x)/f_0(x)$. Although both the density difference and density ratio metrics are intuitive, there are several weaknesses associated with each of them. For example, the estimation of a density difference is largely limited to kernel density estimators, which are sensitive to the curse of dimensionality. The density ratio, on the other hand, could potentially be estimated

using various regression methods thanks to the following reformulation:

$$\frac{f_1(x)}{f_0(x)} = \frac{\pi_0}{\pi_1}\frac{m(x)}{1-m(x)},$$

(see, e.g., Qin and Zhang, 1997). The main weakness of the ratio approach, however, is that the ratio is highly sensitive to the tail behavior of distributions, and it is not even well defined when $m(x) = 1$. To overcome these limitations, we propose an alternative approach which instead compares the regression function with the class probability. More specifically, we consider

$$\mathcal{T}_{global} = \int_S \{m(x) - \pi_1\}^2 dP_X(x), \quad \mathcal{T}_{local}(x) = \{m(x) - \pi_1\}^2 \tag{3.2}$$

as global and local measures of the discrepancy between two distributions where we assume that π_1 is a fixed constant within $0 < \pi_1 < 1$ throughout this chapter. By construction, both $\mathcal{T}_{global}$ and $\mathcal{T}_{local}(x)$ are bounded between zero and one. More importantly, we can take advantage of numerous existing regression methods (see, e.g., Friedman et al., 2009, for popular methods and descriptions) when estimating $m(x)$. Hence, our approach maintains the flexibility of the density ratio approach while avoiding the problem of ill-defined quantities.

3.2.2 Test Statistics and Algorithms

Suppose we observe n pairs of samples $\{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^D$ and $Y_i \in \{0, 1\}$. Let $\widehat{m}(x)$ be an estimate of $m(x)$ based on the samples, and $\widehat{\pi}_1 = \frac{1}{n}\sum_{i=1}^n I(Y_i = 1)$. Then by plugging these statistics into (3.2), we define our global and local test statistics as

$$\widehat{\mathcal{T}}_{global} = \frac{1}{n}\sum_{i=1}^n \{\widehat{m}(X_i) - \widehat{\pi}_1\}^2, \quad \widehat{\mathcal{T}}_{local}(x) = \{\widehat{m}(x) - \widehat{\pi}_1\}^2. \tag{3.3}$$

The null distributions of the proposed test statistics are typically unknown, and they depend on the choice of regression method as well as the distribution of the data. Hence, to keep our framework as general as possible, we use a permutation procedure to set a critical value that yields a valid level α test for any given regression estimator under any sampling scheme given in Section 3.2.3. The proposed permutation framework for global and local two-sample testing are summarized in Algorithm 1 and Algorithm 2, respectively.

3.2.3 Sampling Schemes

In the two-sample problem, there are two common sampling schemes for obtaining the paired data set $\{(X_i, Y_i)\}_{i=1}^n$, namely i) *i.i.d. sampling* and ii) *separate sampling* defined as follows:

Algorithm 1: Global Two-Sample Testing via Permutations

Require: samples $\{X_i, Y_i\}_{i=1}^n$, number of permutations B, significance level α, a regression method.

(1) Calculate the global test statistic $\widehat{\mathcal{T}}_{global}$.

(2) Randomly permute $\{Y_1, \ldots, Y_n\}$. Calculate the test statistic using the permuted data.

(3) Repeat the previous step B times to obtain $\{\widehat{\mathcal{T}}_{global}^{(1)}, \ldots, \widehat{\mathcal{T}}_{global}^{(B)}\}$.

(4) Approximate the permutation p-value by

$$p = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} I(\widehat{\mathcal{T}}_{global}^{(b)} > \widehat{\mathcal{T}}_{global})\right).$$

(5) Reject the null hypothesis when $p < \alpha$. Otherwise, accept the null hypothesis.

Algorithm 2: Local Two-Sample Testing via Permutations

Require: samples $\{X_i, Y_i\}_{i=1}^n$, test points $\{x_j\}_{j=1}^k$, number of permutations B, significance level α, a multiple testing procedure, a regression method.

(1) Calculate the test statistic $\widehat{\mathcal{T}}_{local}(x_j)$ at the k test points.

(2) Randomly permute $\{Y_1, \ldots, Y_n\}$. Calculate the test statistic using the permuted data.

(3) Repeat the previous step B times to obtain $\{\widehat{\mathcal{T}}_{local}^{(1)}(x_j)\}_{j=1}^k, \ldots, \{\widehat{\mathcal{T}}_{local}^{(B)}(x_j)\}_{j=1}^k$.

(4) Approximate the permutation p-value at each test point x_j by

$$p_j = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} I(\widehat{\mathcal{T}}_{local}^{(b)}(x_j) > \widehat{\mathcal{T}}_{local}(x_j))\right).$$

(5) Apply a multiple testing procedure for controlling the FWER or the FDR at α level.

(6) Return the significant local test points.

- **i.i.d. sampling.** Under i.i.d. sampling, we observe n pairs of i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$ from the joint distribution $\pi_1[P_1 \times \delta_1] + \pi_0[P_0 \times \delta_0]$. Here we note that n is fixed in advance. Then $n_1 = \sum_{i=1}^n I(Y_i = 1)$ and $n_0 = n - n_1$ are Binomial(n, π_1) and Binomial(n, π_0), respectively. This setting is common in applications of supervised learning where the goal is to build a model that can successfully predict the class label Y given the feature vector X (e.g. Friedman et al., 2009). Our goal, on the other hand, is to test whether the two distributions P_0 and P_1 are the same or not by leveraging existing methods in the regression literature.

- **Separate sampling.** In the case of separate sampling, n_0 and n_1 are predetermined and they are not random. We then observe n_0 and n_1 independent sample points from P_0 and P_1 separately, which provides the data set $\{(X_i, Y_i)\}_{i=1}^n$ where $Y_i = 1$ if X_i was drawn from P_1 and $Y_i = 0$ otherwise.

We can link the separate sampling to the i.i.d. sampling scheme by randomly ordering the (X_i, Y_i) pairs, so that the data points are exchangeable and for each $i \in \{1, \ldots, n\}$, the conditional distribution of Y_i given $X_i = x$ is $m(x) = \pi_1 f_1(x)/\{\pi_1 f_1(x) + \pi_0 f_0(x)\}$ where the class probability is given by $\pi_1 = n_1/n$. Therefore,

although the joint distributions of $\{(X_i, Y_i)\}_{i=1}^n$ are different under i.i.d. and separate sampling schemes, they share the same regression function.

Remark 3.1. *These two sampling schemes are also known as* prospective sampling *and* retrospective (or case-control) sampling*, respectively, and their relationships have been studied in different contexts. For example, it has been shown that the logistic slope estimates have similar behaviors under both sampling schemes (see, e.g. Anderson, 1972; Prentice and Pyke, 1979; Wang and Carroll, 1993, 1999; Bunea and Barbu, 2009). This result has been extended to general regression models by Scott and Wild (2001).*

3.3 Global Two-Sample Tests via Regression

The choice of regression method in our framework will ultimately decide whether we achieve competitive statistical power. In Section 3.3.1, we illustrate the point that the global regression test can be optimal if we choose a suitable regression method. For this theoretical purpose, we focus on the regression test based on Fisher's LDA and show its optimality. In Section 3.3.2, we turn our attention to more general regression settings and characterize the testing error of the global regression test in terms of the mean integrated square error (MISE) of the regression estimator.

3.3.1 Fisher's Linear Discriminant Analysis

In this section, we consider a simple scenario of two sample normal mean to highlight the difference between our approach and the classification accuracy approach. In particular, we prove that the regression test based on Fisher's LDA achieves the same local power as Hotelling's T^2 test. This result has significance given that i) Hotelling's test is optimal under the considered scenario and ii) the classification accuracy test based on Fisher's LDA is usually underpowered (Ramdas et al., 2016; Rosenblatt et al., 2016). To facilitate comparison with the previous results, which are established under separate sampling, we also consider the case where n_0 and n_1 are predetermined throughout this subsection.

Suppose we observe $\{X_{i,0}\}_{i=1}^{n_0} \overset{i.i.d.}{\sim} N(\mu_0, \Sigma)$ and independently $\{X_{i,1}\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} N(\mu_1, \Sigma)$. We denote the pooled samples by $\{X_i\}_{i=1}^n = \{X_{i,0}\}_{i=1}^{n_0} \cup \{X_{i,1}\}_{i=1}^{n_1}$ where $n = n_0 + n_1$. The two-sample problem then becomes the problem of testing for mean differences as

$$H_0 : \mu_0 = \mu_1 \quad \text{versus} \quad H_1 : \mu_0 \neq \mu_1. \tag{3.4}$$

For this particular problem, Fisher's LDA is a natural choice for regression, assuming normality and equal class covariances. Let $\widehat{\mu}_i$ be the sample mean vector for each group, $\mathcal{S}$ be the covariance matrix of the combined samples, i.e. $\mathcal{S} = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$ where $\widehat{\mu} = n^{-1}\sum_{i=1}^n X_i$. Then, by putting

$\pi_1 = n_1/n$, the regression estimator based on Fisher's LDA is given by

$$\begin{aligned} & \widehat{m}_{\mathrm{LDA}}(x) \\ = \; & \frac{\pi_1 \exp\big\{-\frac{1}{2}(x-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(x-\widehat{\mu}_1)\big\}}{\pi_0 \exp\big\{-\frac{1}{2}(x-\widehat{\mu}_0)^\top \mathcal{S}^{-1}(x-\widehat{\mu}_0)\big\} + \pi_1 \exp\big\{-\frac{1}{2}(x-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(x-\widehat{\mu}_1)\big\}}. \end{aligned} \tag{3.5}$$

One of the most popular test statistics for testing (3.4) is Hotelling's T^2 statistic, which yields optimal power for the normal means problem (see, e.g. Anderson, 2003). For the two-sample problem, Hotelling's T^2 statistic is defined by

$$T^2_{\mathrm{Hotelling}} = \frac{n_0 n_1}{n_0+n_1}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}_p^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1),$$

where $\mathcal{S}_p$ is the pooled covariance matrix, that is

$$\mathcal{S}_p = \frac{1}{n_0+n_1-2}\left(\sum_{i=1}^{n_0}(X_{i,0}-\widehat{\mu}_0)(X_{i,0}-\widehat{\mu}_0)^\top + \sum_{i=1}^{n_1}(X_{i,1}-\widehat{\mu}_1)(X_{i,1}-\widehat{\mu}_1)^\top\right).$$

On the other hand, the regression test statistic based on Fisher's LDA is given by

$$\widehat{\mathcal{T}}_{\mathrm{LDA}} = \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{m}_{\mathrm{LDA}}(X_i) - \pi_1\Big)^2.$$

The next theorem provides a connection between the seemingly unrelated $\widehat{\mathcal{T}}_{\mathrm{LDA}}$ and $T^2_{\mathrm{Hotelling}}$ statistics. Specifically, it shows that $n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\mathrm{LDA}}$ is asymptotically identical to Hotelling's T^2 statistic under the null. It is also worth pointing out that the theorem still holds without the normality assumption.

Theorem 3.1. *Let $\{X_{i,0}\}_{i=1}^{n_0}$ and $\{X_{i,1}\}_{i=1}^{n_1}$ be random samples under separate sampling from two multivariate distribution with the mean vectors μ_0 and μ_1, respectively, and the same covariance matrix Σ. Assume the pooled samples are mutually independent and the third moments of $X_{1,0}$ and $X_{1,1}$ are finite. Suppose that $\mathcal{S}_p$ and $\mathcal{S}$ satisfy $\mathcal{S}_p^{-1} = \Sigma^{-1}(1+o_P(1))$ and $\mathcal{S}^{-1} = \Sigma^{-1}(1+o_P(1))$. Then, under $H_0 : \mu_0 = \mu_1$, it holds that*

$$n\widehat{\mathcal{T}}_{\mathrm{LDA}} = n\pi_0^2\pi_1^2(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}_p^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) + o_P(1). \tag{3.6}$$

Therefore,

$$n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\mathrm{LDA}} = T^2_{\mathrm{Hotelling}} + o_P(1) \xrightarrow{d} \chi^2_D,$$

where χ^2_D is the chi-squared distribution with D degrees of freedom.

Let us now turn to the alternative hypothesis. To begin with, we consider a family of probability functions that satisfy the following smoothness condition.

Definition 3.1 (Definition 12.2.1 of Lehmann and Romano (2006)). *Let $\{P_\mu, \mu \in \Omega\}$ be a parametric model where Ω is an open subset of $\mathbb{R}^D$, and let $f_\mu(x) = dP_\mu(x)/d\nu(x)$ be the density function with respect to Lebesgue measure ν. The family $\{P_\mu, \mu \in \Omega\}$ is quadratic mean differentiable (q.m.d.) at μ_0 if there exists a vector of real-valued functions $\eta(\cdot, \mu_0) = (\eta_1(\cdot, \mu_0), \cdots, \eta_D(\cdot, \mu_0))^\top$ such that*

$$\int_{\mathbb{R}^D} \left[\sqrt{f_{\mu_0+h}(x)} - \sqrt{f_{\mu_0}(x)} - \langle \eta(x, \mu_0), h \rangle \right]^2 d\nu(x) = o(||h||_2^2) \tag{3.7}$$

as $||h||_2 \to 0$.

Such q.m.d. families include fairly large parametric models such as exponential families in natural form. For our purpose, we focus on location q.m.d. families, denoted by $\{\mathbb{P}_\mu, \mu \in \Omega\}$. Specifically, $\mathbb{P}_\mu$ is a member of $\{\mathbb{P}_\mu, \mu \in \Omega\}$ if its density satisfies $f_\mu(x) = f(x - \mu)$ for which $f(x)$ has zero mean and covariance matrix Σ. Next, for given $\mathbb{P}_{\mu_0}$ and $\mathbb{P}_{\mu_1}$ from $\{\mathbb{P}_\mu, \mu \in \Omega\}$, let us consider the local alternative

$$H_{1,n} : \mu_1 - \mu_0 = h/\sqrt{n}, \tag{3.8}$$

where $h = (h_1, \ldots, h_D)^\top$. Then, under $H_{1,n}$, $\widehat{\mathcal{T}}_{\text{LDA}}$ has asymptotic behavior as follows.

Theorem 3.2. *Suppose under separate sampling that $\{X_{i,0}\}_{i=1}^{n_0} \overset{i.i.d.}{\sim} \mathbb{P}_{\mu_0}$ and independently $\{X_{i,1}\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} \mathbb{P}_{\mu_1}$ where $\mathbb{P}_{\mu_i}$ is a member of the location q.m.d. family with the same covariance matrix Σ and finite third moments. Suppose that $\mathcal{S}_p$ and $\mathcal{S}$ satisfy $\mathcal{S}_p^{-1} = \Sigma^{-1}(1 + o_P(1))$ and $\mathcal{S}^{-1} = \Sigma^{-1}(1 + o_P(1))$. Under the sequence of local alternatives given in (3.8), we have*

$$n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\text{LDA}} = T^2_{\text{Hotelling}} + o_P(1) \xrightarrow{d} \chi^2_D(\lambda),$$

where $\chi^2_D(\lambda)$ denotes a noncentral chi-square distribution with D degrees of freedom and the noncentral parameter

$$\lambda = \pi_0\pi_1 h^\top \Sigma^{-1} h.$$

The results from Theorem 3.1 and Theorem 3.2 imply that our regression test based on $\widehat{\mathcal{T}}_{\text{LDA}}$ has the same asymptotic local power as Hotelling's T^2 test. As a result, the regression test based on $\widehat{\mathcal{T}}_{\text{LDA}}$ is asymptotically optimal against the local alternatives as Hotelling's T^2 test.

To illustrate the main point of this section, we compare the performance of $\widehat{\mathcal{T}}_{\text{LDA}}$ with Hotelling's T^2 test through Monte Carlo simulations. We randomly generate $n_0 = n_1 = 100$ samples from $N((0, \ldots, 0)^\top, I_D)$ and $N((\mu, \ldots, \mu)^\top, I_D)$, respectively and set $\mu^2 = 0.05$ for $D = 5$ and $\mu^2 = 0.01$ for $D = 20$. We also consider

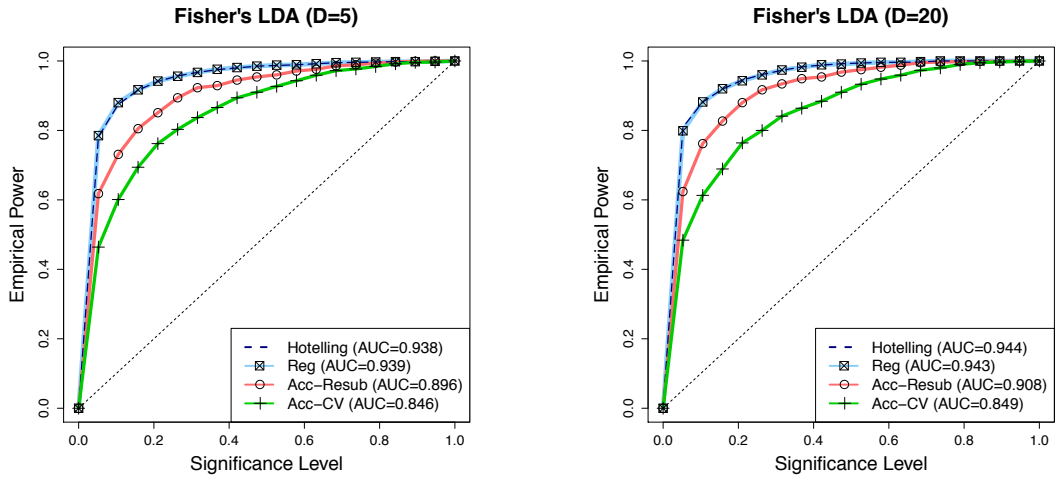

Figure 3.2: Power comparisons between Hotelling's T^2 (Hotelling), $\widehat{\mathcal{T}}_{\text{LDA}}$ (Reg), the in-sample accuracy (Acc-Resub), and the cross-validated accuracy (Acc-CV) via Fisher's LDA.

two versions of the accuracy-based tests via Fisher's LDA: the in-sample (re-substitution) accuracy and the two-fold cross-validated accuracy. To calculate the cross-validated accuracy, we use the balanced sample splitting scheme in which the first part of data is used to train the LDA, and the second part is used to estimate the accuracy of the classifier (see, Definition 1 and 2 of Rosenblatt et al., 2016, for more details). To make a fair comparison, the critical values of the given tests were all decided by the permutation procedure. As shown in Figure 3.2, the regression test based on $\widehat{\mathcal{T}}_{\text{LDA}}$ has comparable power to Hotelling's T^2 test that coincides with our theory. On the other hand, the accuracy tests have less power than Hotelling's T^2 test.

3.3.2 The MISE and Testing Error for Global Regression

We now turn to more general regression settings and investigate general properties of the global regression test in both separate and i.i.d. sampling cases. Let $\mathcal{M}$ be a certain class of regression $m(x) : S \subseteq \mathbb{R}^D \mapsto [0,1]$ containing constant functions. Suppose that we have a regression estimator $\widehat{m}(x)$ that has the mean integrated square error as

$$\sup_{m\in\mathcal{M}} \mathbb{E}\int_S (\widehat{m}(x) - m(x))^2 \, dP_X(x) \leq C_0\delta_n, \tag{3.9}$$

where C_0 is a positive constant and $\delta_n = o(1)$. In the case of i.i.d. sampling, we further assume $\delta_n \geq n^{-1}$, which is typical for nonparametric regression estimators. Our main interest here is in employing the above MISE to characterize the testing error of the global regression test. Note that the plug-in global statistic in (3.3) is typically a biased estimator of the MISE and the bias differs from case to case. To simplify our analysis, we consider sample splitting where the half of data is used to estimate the regression function and

the other is used to evaluate the empirical squared error. In detail, given samples $(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})$, the regression test statistic based on (random) sample splitting is defined by

$$\widehat{\mathcal{T}}'_{global} = \frac{1}{n} \sum_{i=n+1}^{2n} \left(\widehat{m}(X_i) - \widehat{\pi}_1\right)^2, \tag{3.10}$$

where $\widehat{m}(\cdot)$ and $\widehat{\pi}_1$ are calculated based on the first half of the data $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. In the case of separate sampling, we assume a random ordering in the entire data set and similarly split it into two parts but with the additional restriction that class probabilities are the same in both parts. Based on $\widehat{\mathcal{T}}'_{global}$, we argue that for sufficiently large $C_1 > 0$ and n, the testing error of the global regression test can be arbitrarily small against the class of global alternatives given by

$$\mathcal{M}(C_1\delta_n) = \left\{ m \in \mathcal{M} : \int_S \left(m(x) - \pi_1\right)^2 dP_X(x) \geq C_1\delta_n \right\}.$$

Note that since π_1 is assumed to be fixed, the regression function $m(x)$ is completely determined by f_0 and f_1. Thus in the following theorem and hereafter, we use the notation $f_0, f_1 \in \mathcal{M}$ to represent $m(x) = \pi_1 f_1(x)/\{\pi_0 f_0(x) + \pi_1 f_1(x)\} \in \mathcal{M}$. Similarly, we write $f_0, f_1 \in \mathcal{M}_0$ to signify that $\pi_1 f_1(x)/\{\pi_0 f_0(x) + \pi_1 f_1(x)\} = \pi_1$ for all $x \in S$. With this notation in hand, we state the main theorem of this subsection.

Theorem 3.3. *Consider the case of i.i.d. sampling or separate sampling. In each case, suppose that we have a regression estimator $\widehat{m}(\cdot)$ satisfying (3.9). Let t_α be the upper α quantile of the permutation distribution of $\widehat{\mathcal{T}}'_{global}$ based on $\widehat{m}(\cdot)$ where we permute the first half of labels. For fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, we assume that there exists a positive constant $C'_{0,\alpha}$ such that $\sup_{f_0,f_1\in\mathcal{M}} \mathbb{P}_{f_0,f_1}(t_\alpha < C'_{0,\alpha}\delta_n) \geq 1 - \beta/2$. Then there exist positive constants C_1 and N depending on $C_0, C'_{0,\alpha}, \alpha, \beta$ such that*

- *Type I error:* $\sup_{f_0,f_1\in\mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}'_{global} > t_\alpha\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n\geq N} \sup_{f_0,f_1\in\mathcal{M}(C_1\delta_n)} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}'_{global} \leq t_\alpha\right) \leq \beta.$

Theorem 3.3 uses the assumption that the permutation critical value of the regression test is uniformly bounded by δ_n (up to some constant factor) with high probability. We end this subsection with a class of regression estimators, which satisfy this assumption. Let us consider a class of regression estimators with the following representation:

$$\widehat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

where $w_i(x) \geq 0$ and $\sum_{i=1}^n w_i(x) = 1$ for all x. In addition, we assume that $w_i(x)$ is a function of $\{X_1, \ldots, X_n\}$ but not $\{Y_1, \ldots, Y_n\}$. This class of estimators, often called linear smoothers, contains many popular regression methods such as k-nearest neighbor (kNN) regression, kernel regression and local polynomial regression. Focusing on linear smoothers, we provide the following corollary.

Corollary 3.3.1. *Consider the case of i.i.d. sampling or separate sampling. In each case, let $\widehat{\mathcal{T}}'_{global}$ be the global regression test statistic in (3.10) based on a linear smoother $\widehat{m}(\cdot)$ with the property in (3.9). Let t_α be the upper α quantile of the permutation distribution of $\widehat{\mathcal{T}}'_{global}$ where we permute the first half of labels. Then for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exist positive constants C_1 and N depending on C_0, α, β such that*

- *Type I error:* $\sup_{f_0, f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}'_{global} > t_\alpha\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N} \sup_{f_0, f_1 \in \mathcal{M}(C_1 \delta_n)} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}'_{global} \leq t_\alpha\right) \leq \beta.$

3.3.3 Examples

In the case of i.i.d. sampling, the convergence rate δ_n of commonly used regression estimators have been well-established and these results can be directly used to study the testing error of the global regression test. We list several known results here. More examples can be found in Györfi et al. (2002), Tsybakov (2009) and Devroye et al. (2013).

- **kNN regression.** When $\mathcal{M}$ is a class of Lipschitz continuous functions, the convergence rate of kNN estimators satisfies $\delta_n = n^{-2/(2+D)}$ (Györfi et al., 2002). This can be generalized to a Hölder space with smooth parameter β in which the rate becomes $\delta_n = n^{-2\beta/(2\beta+D)}$ (Györfi et al., 2002; Ayano, 2012) for $0 < \beta \leq 1.5$. Furthermore, Kpotufe (2011) shows that kNN estimators are adaptive to the intrinsic dimension $d \ll D$ under appropriate conditions. In this case, the convergence rate becomes much faster as $\delta_n = n^{-2/(2+d)} \ll n^{-2/(2+D)}$.

- **Kernel regression.** Kernel regression estimators also achieve the converge rate as $\delta_n = n^{-2/(2+D)}$ for Lipschitz continuous functions and more generally as $\delta_n = n^{-2\beta/(2\beta+D)}$ for a Hölder space with smooth parameter $0 < \beta \leq 1.5$ (Györfi et al., 2002). The adaptivity of kernel regression to the intrinsic dimension has been proved by Kpotufe and Garg (2013). Following their results, the convergence rate becomes $\delta_n = n^{-2/(2+d)} \ll n^{-2/(2+D)}$ when there exists a low-dimensional structure in the data.

- **Local polynomial regression.** Let $\mathcal{M}$ be a Sobolev space with smoothness α. Then local polynomial regression estimators has the convergence rate as $\delta_n = n^{-\alpha/(\alpha+d)}$ where d is manifold dimension smaller

than the original dimension D (Bickel and Li, 2007).

- **Random forests regression.** For Lipschitz continuous functions, Biau (2012) shows that the random forest estimator converges at rate $\delta_n = n^{-\frac{0.75}{s \log 2 + 0.75}}$ where s is the number of the relevant features. Hence, the convergence rate of the random forests becomes faster than $n^{-2/(2+D)}$ when $s \leq D/2$ under certain conditions. Wager and Walther (2015) use the guess-and-check forest algorithm to show that the convergence rate of the random forest is $\delta_n = n^{-\log(\xi)/\log(2\xi)}$ where $\xi = 1/(1 - 3/4s)$.

To the best of our knowledge, there has been no detailed investigation of the regression estimation error under separate sampling. In this case, we cannot directly take advantage of existing results on regression. However, as the sample size becomes larger, the difference between i.i.d. sampling and separate sampling becomes minor. Hence we expect that a reasonable regression estimator behaves similarly under both sampling schemes in large sample sizes, while a detailed analysis is necessary in future work. It is also worth noting that for certain regression methods, consistency results are not significantly affected by sampling scheme. For example, the consistency theory for L_1 penalized regression relies mainly on the assumption about a design matrix, which can be fulfilled under both sampling schemes (Van de Geer, 2008; Bühlmann and Van De Geer, 2011). In such a case, the same convergence rate can be established under both sampling schemes.

3.4 Local Two-Sample Tests via Regression

The global two-sample test only answers the question whether two distributions are different, whereas in some applications, it would be more valuable to describe how these two distributions differ in a multivariate space. With this goal in mind, we now move on to the local two-sample problem and study general properties of the local regression test.

3.4.1 The MSE and Testing Error for Local Regression

We start by establishing similar results in Section 3.3.2 for local regression tests. Given a local point $x \in S$ of interest, suppose that a regression estimator has the mean square error such that

$$\sup_{m \in \mathcal{M}} \mathbb{E}\left[\left(\widehat{m}(x) - m(x)\right)^2\right] \leq C_{0,x}\delta_{n,x}, \tag{3.11}$$

where $C_{0,x}$ is a positive constant and $\delta_{n,x} = o(1)$. In addition, we assume $\delta_{n,x} \geq n^{-1}$ for i.i.d. sampling. Then the next theorem shows that for sufficiently large $C_{1,x}$ and n, the local testing error based on the given

regression estimator can be arbitrarily small against the class of local alternatives given by

$$\mathcal{M}(C_{1,x}\delta_{n,x}) = \Big\{ m \in \mathcal{M} : (m(x) - \pi_1)^2 \geq C_{1,x}\delta_{n,x} \Big\}.$$

Theorem 3.4. *Consider the case of i.i.d. sampling or separate sampling. In each case, consider the local regression test statistic $\widehat{\mathcal{T}}_{local}(x)$ in (3.3) based on a linear smoother $\widehat{m}(x) = \sum_{i=1}^n w_i(x)Y_i$ with the property in (3.11). Let t_α be the upper α quantile of the permutation distribution of $\widehat{\mathcal{T}}_{local}(x)$. Then for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that*

- *Type I error:* $\sup_{f_0, f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{local}(x) > t_\alpha\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}(C_{1,x}\delta_{n,x})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{local}(x) \leq t_\alpha\right) \leq \beta.$

Remark 3.2. *Although Theorem 3.4 focuses on a linear smoother, the same conclusion holds for other regression methods as long as there exists a positive constant $C_{0,x,\alpha}$ such that the permutation critical value t_α is bounded above by $C_{0,x,\alpha}\delta_n$ with high probability (see Theorem 3.3 for a more formal statement).*

In order to keep things as simple and concrete as possible, we next focus on the Lipschitz class and analyze the optimality of the local regression tests from a minimax point of view. In the rest of this section (Section 3.4.2–3.4.4), we concentrate on *i.i.d. sampling scheme* to take full advantage of known regression results. However, as we discussed in Section 3.3.3, similar results are expected to hold under separate sampling as well.

3.4.2 Minimax Optimality over the Lipschitz Class

For a fixed constant $L > 0$, let us denote the Lipschitz function class by

$$\mathcal{M}_{Lip} = \Big\{ m : |m(x) - m(y)| \leq L||x - y||_2 \text{ for all } x, y \in S \Big\}.$$

We also denote the collection of α level tests by $\Phi_{n,\alpha} = \{\phi : \sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}(\phi = 1) \leq \alpha\}$ and denote the class of Lipschitz local alternatives by

$$\mathcal{M}_{Lip}(\delta_{n,x}) = \Big\{ m \in \mathcal{M}_{Lip} : (m(x) - \pi_1)^2 \geq \delta_{n,x} \Big\}. \tag{3.12}$$

With this notation and fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, the *minimum separation* is defined by

$$\delta^\star_{n,x} = \inf \Big\{ \delta_{n,x} : \inf_{\phi \in \Phi_{n,\alpha}} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(\delta_{n,x})} \mathbb{P}_{f_0,f_1}(\phi = 0) \leq \beta \Big\}, \tag{3.13}$$

which is the smallest distance between $m(x)$ and π_1 such that the power becomes nontrivial. Then a test is called minimax rate optimal if it has power uniformly over $\mathcal{M}_{Lip}(\delta_{n,x})$ such that $\delta_{n,x} \asymp \delta^{\star}_{n,x}$.

In this section, we will investigate minimax rate optimality of local regression tests over the Lipschitz class under i.i.d. sampling. First we formally state an upper bound for the local estimation error based on kNN and kernel regression in Example 3.1 and Example 3.2, respectively. We then use these results to obtain the upper bound for the minimum separation in Corollary 3.4.1.

Example 3.1 (kNN regression)**.** *For a fixed point $x \in S$, list the data by*

$$(X_{1,n}(x), Y_{1,n}(x)), \ldots, (X_{n,n}(x), Y_{n,n}(x)),$$

where $X_{k,n}(x)$ is the kth nearest neighbor of x and $Y_{k,n}(x)$ is its pair. Consider the kNN regression estimator

$$\widehat{m}_{kNN}(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(x), \tag{3.14}$$

and assume that $\mathbb{P}(X \in B_{x,\epsilon}) > \tau_x \epsilon^D$ where $B_{x,\epsilon}$ is a ball of radius $\epsilon > 0$ centered at x and $\tau_x > 0$. Then

$$\sup_{m \in \mathcal{M}_{Lip}} \mathbb{E}\left[\left(\widehat{m}_{kNN}(x) - m(x) \right)^2 \right] \leq \frac{1}{4k_n} + L^2 \frac{2\Gamma(2/D)}{D\tau_x^{2/D}} \left(\frac{k_n}{n} \right)^{2/D},$$

and for $k_n = n^{2/(2+D)}$, we have

$$\sup_{m \in \mathcal{M}_{Lip}} \mathbb{E}\left[\left(\widehat{m}_{kNN}(x) - m(x) \right)^2 \right] \leq C_{0,x} n^{-\frac{2}{2+D}},$$

where $C_{0,x} = 1/4 + L^2 \Gamma(2/D) D^{-1} \tau_x^{-2/D}$.

A similar result can be established for kernel regression estimators as follows.

Example 3.2 (Kernel regression)**.** *Given a kernel $K : S \mapsto [0, \infty)$, the kernel regression estimator at a fixed point x is given by*

$$\widehat{m}_{ker}(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h_n} \right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n} \right)}. \tag{3.15}$$

Assume there exists $0 < r < R$ and $0 < \lambda < 1$ such that

$$\lambda I(x \in B_{0,r}) \leq K(x) \leq I(x \in B_{0,R})$$

where $B_{0,\epsilon}$ is a ball of radius $\epsilon > 0$ centered at the origin. Further assume that $\mathbb{P}(X \in B_{x,\epsilon}) > \tau_x \epsilon^D$ for some $\tau_x > 0$. Then

$$\sup_{m \in \mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{ker}(x) - m(x))^2\right] \leq \left(\frac{1+\lambda}{4\lambda^2 \tau_x r^D} + \frac{2e^{-1}}{\tau_x r^D}\right) \frac{1}{n h_n^D} + L^2 R^2 h_n^2$$

and for $h_n = n^{-2/(2+D)}$,

$$\sup_{f_0, f_1 \in \mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{ker}(x) - m(x))^2\right] \leq C_{0,x} n^{-\frac{2}{2+D}}$$

where $C_{0,x} = (1+\lambda)/(4\lambda^2 \tau_x r^D) + 2e^{-1}/(\tau_x r^D) + L^2 R^2$.

Remark 3.3. *Example 3.1 and Example 3.2 are well-known and standard except that we keep track of the constant $C_{0,x}$ over the Lipschitz class. Similar results exist in the literature but for slightly different settings. Hence, in Appendix C.4, we present detailed proofs for these two examples heavily building on Györfi et al. (2002). The proofs will also be used to study the performance of the kNN and kernel local regression tests under the existence of intrinsic dimension in Proposition 3.1.*

From the previous examples together with Theorem 3.4, we conclude that the minimum separation in (3.13) satisfies $\delta^\star_{n,x} \lesssim n^{-2/(2+D)}$. We summarize this result in the following corollary.

Corollary 3.4.1 (Upper bound)**.** *Let us denote the local kNN and kernel regression test statistics by*

$$\widehat{\mathcal{T}}_{kNN}(x) = (\widehat{m}_{kNN}(x) - \widehat{\pi}_1)^2, \quad \widehat{\mathcal{T}}_{ker}(x) = (\widehat{m}_{ker}(x) - \widehat{\pi}_1)^2, \tag{3.16}$$

and the upper α quantile of the permutation distribution of each statistic by $t_{\alpha,kNN}$ and $t_{\alpha,ker}$ respectively. Suppose the conditions in Example 3.1 holds with $k_n = n^{2/(D+2)}$. Then for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that

- *Type I error:* $\sup_{f_0, f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{kNN}(x) > t_{\alpha,kNN}\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0, f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+D)})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{kNN}(x) \leq t_{\alpha,kNN}\right) \leq \beta.$

On the other hand, under the conditions in Example 3.2 with $h_n = n^{-2/(2+D)}$ and for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that

- *Type I error:* $\sup_{f_0, f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{ker}(x) > t_{\alpha,ker}\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0, f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+D)})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{ker}(x) \leq t_{\alpha,ker}\right) \leq \beta$

As a result, the minimum separation satisfies $\delta^{\star}_{n,x} \lesssim n^{-2/(2+D)}$.

Next based on the standard technique to lower bound the testing error (e.g., Ingster, 1987; Baraud, 2002), we establish a lower bound for the minimum separation by $n^{-2/(2+D)} \lesssim \delta^{\star}_{n,x}$. This results matches with the upper bound in Corollary 3.4.1. Therefore, the tests in Corollary 3.4.1 are minimax rate optimal and cannot be improved.

Theorem 3.5 (Lower bound)**.** *For any given* $\alpha \in (0,1)$ *and* $\beta \in (1-\alpha)$*, there exists a constant* $C_{1,x} > 0$ *such that*

$$\inf_{\phi \in \Phi_{n,\alpha}} \sup_{f_0, f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+D)})} \mathbb{P}_{f_0,f_1}(\phi = 0) \geq \ 1 - \alpha - \beta.$$

Remark 3.4. *In the context of two-sample testing, it is sometimes more natural to make smoothness assumptions on densities* f_0 *and* f_1 *rather than on the regression function. Here we briefly discuss how to translate the smoothness condition on* f_0 *and* f_1 *into a condition on the regression function. Suppose that density functions* f_0 *and* f_1 *are uniformly bounded below by* $c > 0$ *(see, e.g. Yang and Barron, 1999, for a similar assumption). Then some algebra shows that*

$$|m(x) - m(y)| \leq \pi_0 c^{-1} |f_0(x) - f_0(y)| + \pi_1 c^{-1} |f_1(x) - f_1(y)|.$$

In other words, if f_0 *and* f_1 *are Lipschitz continuous (or more generally Hölder continuous), then the regression function is also Lipschitz continuous with a different Lipschitz constant. This means that our theoretical results will remain valid for the class of Lipschitz densities with the boundedness condition.*

3.4.3 An Approach to Intrinsic Dimension

The previous results show that no test is uniformly powerful when the square distance between $m(x)$ and π_1 is order of $n^{-2/(2+D)}$; therefore it demonstrates the typical curse of dimensionality. Suppose that data $X \in S \subseteq \mathbb{R}^D$ has low intrinsic dimension d which is smaller than the original dimension D (e.g. manifold data). In this case, we would like to have a test whose performance only depends on intrinsic dimension and thus avoids the curse of dimensionality. For this purpose, we consider the homogeneous measure which captures local dimension of data.

Definition 3.2. *(Definition 2 of Kpotufe, 2011) Fix* $x \in S \subseteq \mathbb{R}^D$*, and* $r > 0$*. Let* $C > 0$ *and* $1 \leq d < D$*. The probability measure* $\mathbb{P}(\cdot)$ *is* (C,d)*-homogeneous on* $B_{x,r}$ *if we have* $\mathbb{P}(X \in B_{x,r'}) \leq C\epsilon^{-d}\mathbb{P}(X \in B_{x,\epsilon r'})$ *for all* $r' \leq r$ *and* $0 < \epsilon < 1$.

Using Definition 3.2, we reproduce Corollary 3.4.1 and show that the performances of the local kNN and kernel regression tests depend on the intrinsic dimension instead of the original dimension.

Proposition 3.1. *Consider the same notations as in Corollary 3.4.1 and let $x \in S \subseteq \mathbb{R}^D$. Suppose the probability measure $\mathbb{P}(\cdot)$ is (C,d)-homogeneous on $B_{x,r}$. Then for the kNN regression test with $k_n = n^{2/(2+d)}$ and for any $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that*

- *Type I error:* $\sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{kNN}(x) > t_{\alpha,kNN}\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+d)})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{kNN}(x) \leq t_{\alpha,kNN}\right) \leq \beta.$

On the other hand, for the kernel regression test with $h_n = n^{-2/(2+d)}$ and for any $\beta \in (0, 1-\alpha)$, there exist positive constants $C_{1,x}$ and N_x such that

- *Type I error:* $\sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{ker}(x) > t_{\alpha,ker}\right) \leq \alpha$ *and*
- *Type II error:* $\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+d)})} \mathbb{P}_{f_0,f_1}\left(\widehat{\mathcal{T}}_{ker}(x) \leq t_{\alpha,ker}\right) \leq \beta.$

When the intrinsic dimension is unknown, one can employ a Bonferroni procedure to obtain the same results in Proposition 3.1. To illustrate the idea, let $k_n(i) = n^{-2/(i+2)}$ for $i = 1, \ldots, D$ and denote the resulting kNN tests by $\phi_i(\alpha) = I(\mathcal{T}^{(i)}_{kNN}(x) > t^{(i)}_{\alpha,kNN})$ where $\mathcal{T}^{(i)}_{kNN}(x)$ and $t^{(i)}_{\alpha,kNN}$ are the kNN test statistic calculated with $k_n(i)$ and the corresponding α level permutation critical value, respectively. Then the final test is defined by $\phi_{max} = \max_{1 \leq i \leq D} \phi_i(\alpha/D)$. By using the union bound, it is easy to see that $\sup_{f_0,f_1 \in \mathcal{M}_0} \mathbb{P}_{f_0,f_1}(\phi_{max} = 1) \leq \alpha$ and

$$\sup_{n \geq N_x} \sup_{f_0,f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+d)})} \mathbb{P}_{f_0,f_1}(\phi_{max} = 0) \leq \beta,$$

for certain $C_{1,x}$ and N_x. This shows that the Bonferroni test does not lose any power in terms of separation rate and it adapts to the unknown intrinsic dimension. Despite this theoretical guarantee, the Bonferroni approach should be used with caution in practice. Indeed the Bonferroni test might be too conservative since it does not take into account the dependency structure among $\phi_1, \ldots, \phi_D$.

Remark 3.5. *For simplicity, we illustrate our idea on the Lipschitz class which only requires a mild smoothness assumption. Nevertheless our results in Section 3.4.2–3.4.3 can be extended to a general function class such as Hölder class (e.g. Chapter 3.2 of Györfi et al., 2002) in a similar way. Indeed, all we need is a uniform bound for the MSE (3.11) over a general class, which can be found in the regression literature (See Section 3.3.3).*

3.4.4 Limiting Distribution of Local Permutation Test Statistics

When the sample size is large, calculating the permutation distribution is time-consuming. Hence it would be useful to investigate the limiting distribution of the permutation statistic. Based on the combinatorial central limit theorem (e.g. Bolthausen, 1984), we show that the permutation distribution of our local test statistic converges to the chi-square distribution with one degree of freedom as the sample size tends to infinity.

Theorem 3.6. *Consider the local regression test statistic* $\widehat{\mathcal{T}}_{local}(x)$ *in (3.3) based on a linear smoother* $\widehat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i$. *Suppose that*

$$\frac{\max_{1\le i\le n} |w_i(x) - 1/n|}{\{\sum_{i=1}^{n}(w_i(x) - 1/n)^2\}^{1/2}} \xrightarrow{p} 0 \tag{3.17}$$

holds and let

$$\sigma_n^2 = \frac{n}{n-1}\widehat{\pi}_1(1-\widehat{\pi}_1)\sum_{i=1}^{n}\left(w_i(x) - \frac{1}{n}\right)^2. \tag{3.18}$$

Further let $\eta = (\eta_1, \ldots, \eta_n)$ *be a permutation of* $\{1, \ldots, n\}$. *Then the permutation distribution of the one-side local regression statistic converges to the standard normal distribution as*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}_\eta \left(\sigma_n^{-1}(\widehat{m}_\eta(x) - \widehat{\pi}_1) \le t \middle| \mathcal{X}_n \right) - \mathbb{P}\left(N(0,1) \le t\right) \right| \xrightarrow{p} 0.$$

Here $\mathbb{P}_\eta(\cdot|\mathcal{X}_n)$ *is the uniform probability measure over permutations conditioned on* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *and* $\widehat{m}_\eta(x) = \sum_{i=1}^{n} w_i(x) Y_{\eta_i}$. *Thereby,* $\sigma_n^{-2}\widehat{\mathcal{T}}_{local}(x)$ *converges to the chi-square distribution with one degree of freedom as*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}_\eta \left(\sigma_n^{-2}\widehat{\mathcal{T}}_{local}(x) \le t \middle| \mathcal{X}_n \right) - \mathbb{P}\left(\chi_1^2 \le t\right) \right| \xrightarrow{p} 0.$$

We illustrate Theorem 3.6 using kNN and kernel regression and show that both $\sigma_n^{-2}\widehat{\mathcal{T}}_{kNN}(x)$ and $\sigma_n^{-2}\widehat{\mathcal{T}}_{ker}(x)$ converge to the chi-square distribution with one degree of freedom under appropriate conditions.

Corollary 3.6.1 (kNN regression)**.** *Consider the kNN estimator in (3.14) with*

$$\sigma_n^2 = \widehat{\pi}_1(1-\widehat{\pi}_1)\frac{(n-1)(n-k)}{n^2 k}.$$

Then the permutation distribution of $\sigma_n^{-2}\widehat{\mathcal{T}}_{kNN}(x)$ *converges to the chi-square distribution with one degree of freedom when* $n, k \to \infty$ *and* $2k < n$.

Corollary 3.6.2 (Kernel regression)**.** *Consider the kernel regression estimator in (3.15) and assume that* $\sup_t |K(t)| = \mathcal{K} < \infty$, $\int K^2(t)dt < \infty$ *and* $\int K_h(t)dx = 1$ *where* $K_h(t) = h^{-D}K(t/h)$. *Denote the density function of* X *by* $f(\cdot)$. *Assume that* $0 < f(x) < \infty$ *and* $f(\cdot)$ *is twice differentiable at* x. *Further assume that* $nh^D \to \infty$ *and* $h \to 0$. *Then the permutation distribution of* $\sigma_n^{-2}\widehat{\mathcal{T}}_{ker}(x)$ *converges to the chi-square distribution with one degree of freedom where* σ_n^2 *is given in (3.18).*

3.5 Simulations

In this section, we carry out simulation studies for global and local two-sample tests to examine the empirical performance of the proposed methods. Throughout our simulations, we focus on the separate sampling scenarios under which other existing two-sample tests are usually investigated. We begin by comparing the regression test based on random forests (Breiman, 2001) with other benchmark competitors in Section 3.5.1. Next in Section 3.5.2, we illustrate by an example that the classification accuracy tests can fail due to their discrete nature while the corresponding regression tests perform well. We also provide simulation results for the local regression test in Section 3.5.3 to validate our approach.

3.5.1 Random Forests Two-Sample Testing

Random forests have been proven to be a powerful tool for regression and classification problems in many application areas (see e.g., Hamza and Larocque, 2005; Díaz-Uriarte and De Andres, 2006; Cutler et al., 2007; Chen and Ishwaran, 2012). Despite the good performance of random forests in classification and regression problems, only a few works have applied these methods to statistical inference problems. To the best of our knowledge, only Gagnon-Bartsch and Shem-Tov (2019) and Hediger et al. (2019) use random forests for the two-sample problem. Now whereas Gagnon-Bartsch and Shem-Tov (2019) and Hediger et al. (2019) consider an accuracy test based on random forests, we propose a regression test based on random forests. The corresponding test statistic is given by

$$\widehat{\mathcal{T}}_{RF} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{RF}(X_i) - \widehat{\pi}_1\right)^2, \tag{3.19}$$

where $\widehat{m}_{RF}$ is the regression estimator from the random forest algorithm. For our simulation study, we implement both the RF accuracy and regression tests with the `randomForest` package (version 4.6-12) in `R` with default options for the parameters. We found in our simulation study that the in-sample classification accuracy of random forests is typically one even under the null case; therefore, the resulting test has no power against any alternative. For this reason, we instead estimate the classification accuracy from out-of-bag samples (which is a default option provided by the `randomForest` package). Throughout this section, we denote the accuracy test statistic based on random forests by $\widehat{\mathcal{A}}_{RF}$.

Simulation Setting

Our simulations analyze two main settings. The first setting includes dense alternatives where the two distributions are different over a number of coordinates. The second setting, on the other hand, considers sparse alternatives where the two distributions differ in only a few coordinates. We carry out the simulations via the permutation procedure with 100 random permutations, repeated 300 times for all test statistics. The significance level is controlled at $\alpha = 0.05$.

Dense Alternatives. For the dense alternatives, we draw random samples of size $n_0 = n_1 = 20$ and dimension $D = 5, 20, 50, 100, 150$ and 200 from either multivariate normal distributions $N(\mu, \Sigma)$ or multivariate Cauchy distribution $\mathsf{C}(\mu, \Sigma)$ with different location μ and scale Σ parameters. We consider the following scenarios:

- **Dense Normal Location.** Test $N(0, I_D)$ versus $N(\mu, I_D)$, where $\mu = (0.2, 0.2, \ldots, 0.2)^\top$.
- **Dense Cauchy Location.** Test $\mathsf{C}(0, I_D)$ versus $\mathsf{C}(\mu, I_D)$, where $\mu = (0.3, 0.3, \ldots, 0.3)^\top$.
- **Dense Normal Scale.** Test $N(0, I_D)$ versus $N(0, J_D)$, where J_D is a diagonal matrix whose diagonal elements are $(0.6, 0.6, \ldots, 0.6)^\top$.
- **Dense Cauchy Scale.** Test $\mathsf{C}(0, I_D)$ versus $\mathsf{C}(0, J_D)$, where J_D is a diagonal matrix whose diagonal elements are $(0.5, 0.5, \ldots, 0.5)^\top$.

Sparse Alternatives. Similarly, we generate random samples with $n_0 = n_1 = 20$ and $D = 20, 50, 100, 200, 300$ and 400 from either multivariate normal distributions or multivariate Cauchy distributions. We consider the following problems:

- **Sparse Normal Location.** Test $N(0, I_D)$ versus $N(\mu, I_D)$, where $\mu = (2, 0, \ldots, 0)^\top$.
- **Sparse Cauchy Location.** Test $\mathsf{C}(0, I_D)$ versus $\mathsf{C}(\mu, I_D)$, where $\mu = (3, 0, \ldots, 0)^\top$.
- **Sparse Normal Scale.** Test $N(0, I_D)$ versus $N(0, J_D)$, where J_D is a diagonal matrix with diagonal elements $(0.01, 1, \ldots, 1)^\top$.
- **Sparse Cauchy Scale.** Test $\mathsf{C}(0, I_D)$ versus $\mathsf{C}(0, J_D)$, where J_D is a diagonal matrix with diagonal elements $(0.01, 1, \ldots, 1)^\top$.

As a benchmark competitor, we consider the maximum mean discrepancy (MMD) test (Gretton et al., 2012) based on

$$\mathrm{MMD}_n^2 = -\frac{2}{n_0 n_1} \sum_{i,j=1}^{n_0,n_1} k(X_{i,0}, X_{i,1}) + \frac{1}{n_0^2} \sum_{i,j=1}^{n_0} k(X_{i,0}, X_{j,0}) + \frac{1}{n_1^2} \sum_{i,j=1}^{n_1} k(X_{i,0}, X_{j,0}), \tag{3.20}$$

Table 3.1: Power analysis against dense location alternatives at level $\alpha = 0.05$

	Normal Dense Location						*Cauchy Dense Location*					
D	5	20	50	100	150	200	5	20	50	100	150	200
$\widehat{\mathcal{T}}_{\text{RF}}$	0.123	0.187	0.303	0.417	0.573	0.633	**0.157**	**0.370**	**0.607**	**0.803**	**0.893**	**0.950**
$\widehat{\mathcal{A}}_{\text{RF}}$	0.070	0.117	0.233	0.340	0.440	0.510	0.093	0.260	0.503	0.693	0.793	0.857
MMD_n	**0.143**	**0.290**	**0.520**	**0.723**	**0.880**	**0.937**	0.097	0.057	0.053	0.050	0.060	0.040
Energy_n	**0.156**	**0.283**	**0.530**	**0.720**	**0.877**	**0.940**	0.083	0.077	0.073	0.057	0.057	0.057

where $k(x, y)$ is the Gaussian kernel with a bandwidth chosen by the median heuristic, i.e. $k(x, y) = \exp\left(-||x - y||_2^2/\sigma_{\text{median}}\right)$ (see, Gretton et al., 2012, for details). We also consider the Energy test (Székely and Rizzo, 2004; Baringhaus and Franz, 2004) based on

$$\text{Energy}_n = \frac{2}{n_0 n_1} \sum_{i,j=1}^{n_0,n_1} ||X_{i,0} - X_{j,1}||_2 - \frac{1}{n_0^2} \sum_{i,j=1}^{n_0} ||X_{i,0} - X_{j,0}||_2 - \frac{1}{n_1^2} \sum_{i,j=1}^{n_1} ||X_{i,1} - X_{j,1}||_2. \tag{3.21}$$

Simulation Results

Tables 3.1–3.4 summarize our simulation results. We see from Table 3.1 and 3.2 that MMD_n and Energy_n perform better than the regression test ($\widehat{\mathcal{T}}_{RF}$) and the accuracy test ($\widehat{\mathcal{A}}_{RF}$) against the dense normal location and scale alternatives. Indeed, MMD_n and Energy_n are known to be asymptotically optimal against the normal location alternative with the identity covariance matrix (Ramdas et al., 2015). However, they are both moment-based statistics, and hence sensitive to outliers. They are also based on the Euclidean metric. A major issue of the Euclidean and similar metrics is that they assign weights to the coordinates proportional to their scale without screening for irrelevant variables. Consequently, neither MMD_n nor Energy_n can properly deal with sparse alternatives, which explains their poor performance against the sparse location and scale alternatives. On the other hand, the base learner of the random forest algorithm is the decision tree. The usual splitting rule of decision trees is invariant to absolute values (see e.g., Chapter 9.2 of Friedman et al., 2009), which leads to robustness against outliers.

Random forests also have the ability to handle sparse alternatives by randomly selecting a few variables during the tree-growing process. By averaging each tree, random forests eventually put more weight on informative variables. In general, $\widehat{\mathcal{T}}_{RF}$ and $\widehat{\mathcal{A}}_{RF}$ are comparable to or more powerful than MMD_n and Energy_n under the sparse location and scale alternatives. Finally, we note from our simulations that the regression test $\widehat{\mathcal{T}}_{RF}$ exhibits higher power than the accuracy test $\widehat{\mathcal{A}}_{RF}$ for the dense as well as the sparse alternatives.

Table 3.2: Power analysis against dense scale alternatives at level $\alpha = 0.05$

	Normal Dense Scale						*Cauchy Dense Scale*					
D	5	20	50	100	150	200	5	20	50	100	150	200
$\widehat{\mathcal{T}}_{\mathrm{RF}}$	0.133	0.187	0.260	0.350	0.410	0.473	0.287	**0.557**	**0.790**	**0.937**	**0.953**	**0.970**
$\widehat{\mathcal{A}}_{\mathrm{RF}}$	0.097	0.150	0.200	0.277	0.277	0.290	0.230	0.407	0.663	0.783	0.840	0.877
MMD_n	**0.210**	**0.563**	**0.847**	**0.993**	**0.997**	**1.000**	**0.380**	0.380	0.407	0.407	0.400	0.400
Energy_n	0.080	0.263	0.397	0.657	0.847	0.913	0.283	0.293	0.310	0.310	0.313	0.297

Table 3.3: Power analysis against sparse location alternatives at level $\alpha = 0.05$

	Normal Sparse Location						*Cauchy Sparse Location*					
D	20	50	100	200	300	400	20	50	100	200	300	400
$\widehat{\mathcal{T}}_{\mathrm{RF}}$	0.953	0.880	**0.830**	**0.687**	**0.600**	**0.503**	**0.960**	**0.933**	**0.897**	**0.710**	**0.643**	**0.577**
$\widehat{\mathcal{A}}_{\mathrm{RF}}$	0.883	0.817	0.763	0.600	0.523	0.440	0.943	0.877	0.830	0.613	0.540	0.527
MMD_n	**0.977**	**0.943**	0.770	0.587	0.437	0.360	0.147	0.067	0.057	0.043	0.057	0.027
Energy_n	**0.977**	**0.943**	0.770	0.587	0.440	0.367	0.157	0.083	0.043	0.037	0.050	0.040

Table 3.4: Power analysis against sparse scale alternatives at level $\alpha = 0.05$

	Normal Sparse Scale						*Cauchy Sparse Scale*					
D	20	50	100	200	300	400	20	50	100	200	300	400
$\widehat{\mathcal{T}}_{\mathrm{RF}}$	**0.630**	**0.333**	**0.287**	**0.167**	**0.167**	**0.133**	**0.830**	**0.550**	**0.390**	**0.257**	**0.197**	**0.170**
$\widehat{\mathcal{A}}_{\mathrm{RF}}$	0.603	0.297	0.220	0.130	0.120	0.087	0.743	0.467	0.287	0.207	0.170	0.150
MMD_n	0.043	0.057	0.043	0.053	0.060	0.063	0.067	0.033	0.040	0.057	0.063	0.043
Energy_n	0.037	0.050	0.043	0.050	0.060	0.063	0.047	0.047	0.040	0.057	0.053	0.037

3.5.2 A Comparison between Regression and Classification Accuracy Tests

As mentioned earlier, many classifiers are typically estimated by dichotomizing regression estimators. Depending on the alternative, this dichotomization can result in a less powerful accuracy test than the corresponding regression test. We specifically demonstrate this point by considering two commonly used nonparametric regression methods; namely, k-nearest neighbors regression and kernel regression.

Simulation Setting

Recall the kNN estimator and the kernel regression estimator in (3.14) and (3.15), respectively. Using these estimators, the global regression test statistics are given by

$$\widehat{\mathcal{T}}_{kNN} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{kNN}(X_i) - \widehat{\pi}_1\right)^2 \quad \text{and} \quad \widehat{\mathcal{T}}_{ker} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{m}_{ker}(X_i) - \widehat{\pi}_1\right)^2.$$

Here we use the Euclidean distance to measure the pairwise distance between observations for kNN. On the other hand, we consider the Gaussian kernel with a diagonal bandwidth matrix with identical components h for kernel regression. The corresponding accuracy test statistics are

$$\widehat{\mathcal{A}}_{kNN} = \frac{1}{n}\sum_{i=1}^{n} I\Big(I(\widehat{m}_{kNN}(X_i) > 1/2) = Y_i\Big) \quad \text{and}$$

$$\widehat{\mathcal{A}}_{ker} = \frac{1}{n}\sum_{i=1}^{n} I\Big(I(\widehat{m}_{ker}(X_i) > 1/2) = Y_i\Big),$$

respectively. For all tests, we reject the null hypothesis when the test statistic is larger than a permutation critical value.

For the simulation study, we let $\{X_{1,0}, \ldots, X_{n_0,0}\} \overset{i.i.d.}{\sim} N(\mu_0, \sigma_0^2 \times I_D)$ and $\{X_{1,1}, \ldots, X_{1,n_1}\} \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2 \times I_D)$ where $\mu_0 = (0, \ldots, 0)^\top$, $\mu_1 = (0.2, \ldots, 0.2)^\top$, $\sigma_0^2 = 1$, and $\sigma_1^2 = 1.2$. Hence, there exist differences in both the location and scale parameters. We choose the sample sizes $n_0 = n_1 = 50$ and change the dimension from $D = 5$ to $D = 75$ by steps of 10. To compare the performance, we carry out the permutation test with 200 permutations, and the simulations are repeated 1,000 times to estimate the power of the test. We provide results for a range of different values of the tuning parameters: $k = 5, 15, 25$ for the k-NN regression, and $h = 5, 15, 25$ for the kernel regression.

Simulation Results

Simulation results are presented in Figure 3.3 and 3.4. From the results, it is seen that the regression tests consistently outperform the corresponding classification accuracy tests under the given scenario. The power of the accuracy tests even decreases with dimension, whereas the power of the regression tests steadily increases with dimension. The increase in power with dimension is desirable under this scenario because each coordinate presents evidence towards the alternative. The counter-intuitive result for the accuracy tests is due to the fact that the tests employ a dichotomized regression estimator. To explain it more clearly, we borrow some results from Mondal et al. (2015). First, it can be shown by the weak law of large numbers that

$$1)\ D^{-1/2}||X_{i,0} - X_{j,0}||_2 \xrightarrow{p} \sigma_0\sqrt{2} \quad \text{for } 1 \le i < j \le n_0,$$

$$2)\ D^{-1/2}||X_{i,1} - X_{j,1}||_2 \xrightarrow{p} \sigma_1\sqrt{2} \quad \text{for } 1 \le i < j \le n_1,$$

$$3)\ D^{-1/2}||X_{i,0} - X_{j,1}||_2 \xrightarrow{p} \sqrt{\sigma_0^2 + \sigma_1^2 + (\mu_0 - \mu_1)^2}$$

for $1 \le i \le n_0$, $1 \le j \le n_1$, as $D \to \infty$ while n_0 and n_1 are fixed. For the given example, we have $\sigma_0\sqrt{2} < \sqrt{\sigma_0^2 + \sigma_1^2 + (\mu_0 - \mu_1)^2} < \sigma_1\sqrt{2}$, which implies that every instance is closer to an instance from the

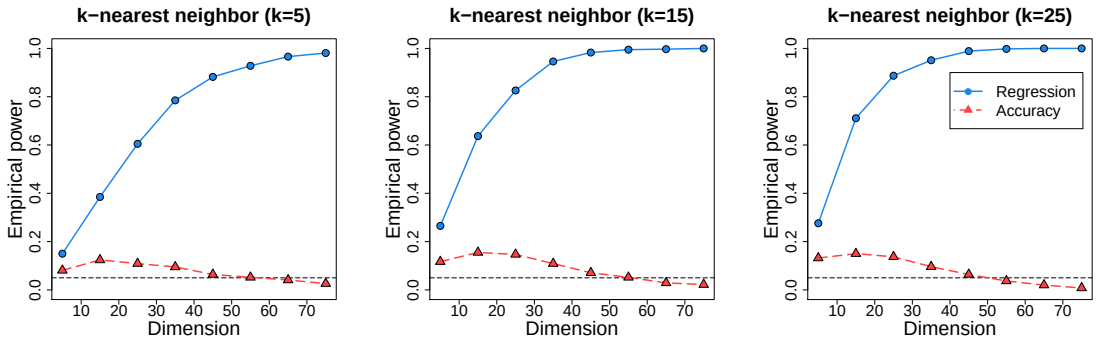

Figure 3.3: Power comparison between the regression test and the classification accuracy test via k-NN regression at level $\alpha = 0.05$ for the toy example in Section 3.5.2.

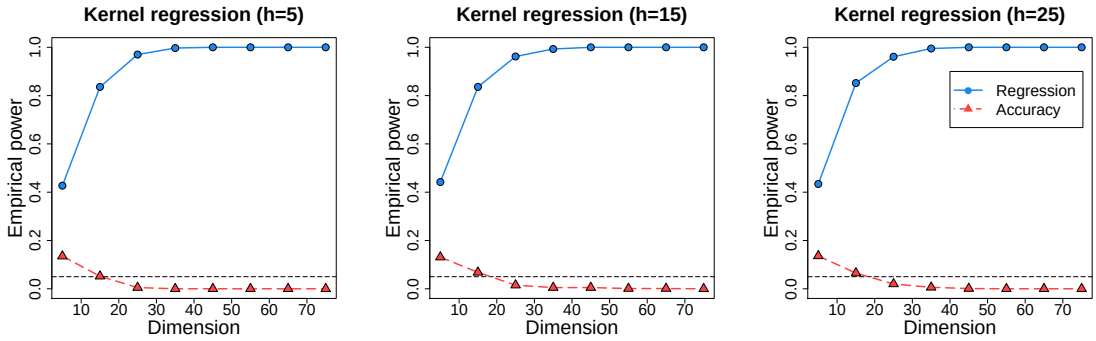

Figure 3.4: Power comparison between the regression test and the classification accuracy test via kernel regression at level $\alpha = 0.05$ for the toy example in Section 3.5.2.

class $Y = 0$ than to other instances from the class $Y = 1$. In other words, the nearest neighbors of any observation are most likely to be from the class $Y = 0$. Note that both k-NN and kernel regression, explicitly or implicitly, use the Euclidean distance to calculate the proximity between two instances. Therefore, we observe with high probability that $\widehat{m}_{kNN}(X_i)$ and $\widehat{m}_{KerR}(X_i)$ are estimated as less than half and the dichotomized classifiers become

$$I\left(\widehat{m}_{kNN}(X_i) > 1/2\right) = \mathcal{I}\left(\widehat{m}_{KerR}(X_i) > 1/2\right) = 0, \quad \text{for all } i = 1, \ldots, n.$$

Due to this dichotomization, $\widehat{\mathcal{A}}_{kNN}$ and $\widehat{\mathcal{A}}_{KerR}$ converge to the empirical class probability n_0/n under the alternative, resulting in poor power performance. On the other hand, the regression tests based on $\widehat{\mathcal{T}}_{kNN}$ and $\widehat{\mathcal{T}}_{ker}$ can be powerful as long as $\widehat{m}_{kNN}(x)$ and $\widehat{m}_{ker}(x)$ significantly deviate from the class probability. This is indeed the case under the considered scenario and thus explains why the regression tests outperform the corresponding classification tests.

3.5.3 Toy Examples for Local Two-Sample Testing

Contrary to classification accuracy, our regression approach naturally leads to a local two-sample testing framework that provides further information on pointwise differences between two populations. We consider two toy examples to demonstrate the empirical performance of the local regression test. For the simulation study, we focus on the local kNN regression statistic in (3.16) with $k_n = n^{2/(2+D)}$ for the normal mixture example and $k_n = n^{2/(2+d)}$ for the manifold example. For both examples, we control the family-wise error rate (FWER) at $\alpha = 0.05$ via the Hochberg step up procedure (Hochberg, 1988).

Normal mixture example

In the first example, we consider two normal mixtures in $\mathbb{R}^2$:

$$f_0(x,y) = \frac{1}{8}\sum_{i=1}^{8}\phi_i(x,y) \quad \text{and} \quad f_1(x,y) = \frac{1}{8}\sum_{i=1}^{8}\phi_i'(x,y),$$

where ϕ_i is the bivariate normal density function with means $\mu_1 = (-3,-3)$, $\mu_2 = (-3,1)$, $\mu_3 = (-1,-1)$, $\mu_4 = (-1,3)$, $\mu_5 = (1,-3)$, $\mu_6 = (1,1)$, $\mu_7 = (3,-1)$, $\mu_8 = (3,3)$ and covariance matrix $\Sigma = 0.3^2 \times I_2$. ϕ_i' is similarly defined with means $\mu_1' = (-3,-1)$, $\mu_2' = (-3,3)$, $\mu_3' = (-1,-3)$, $\mu_4' = (-1,1)$, $\mu_5' = (1,-1)$, $\mu_6' = (1,3)$, $\mu_7' = (3,-3)$, $\mu_8' = (3,1)$ and the same covariance matrix. We generated $n_0 = n_1 = 2000$ samples from f_0 and f_1 and implemented Algorithm 2 to capture local significant points. The local tests were performed at a fixed uniform grid of 50×50 points over $(x,y) \in [-4,4] \times [-4,4]$ and the result is presented in Figure 3.5.

Manifold data example

In the second example, we create high-dimensional data with a low-dimensional manifold structure by generating edge images of size 16×16. Let x, y be integers on evenly spaced points between -30 and 30 that are 2 units apart. Hence the size of the domain of (x,y) becomes 16×16. Given two underlying parameters $\theta \in [-\pi, \pi]$ and $\rho \in [-5,5]$, an edge image is defined by

$$\mathcal{I}(x,y) = I\left(x \cdot \cos(\theta) + y \cdot \sin(\theta) - \rho > 0\right).$$

For the simulation, we draw $n_0 = n_1 = 100$ samples from

$$(\theta_0, \rho_0) \sim \frac{1}{10}\text{Unif}([0,\pi] \times [0,5]) + \frac{9}{10}\text{Unif}([-\pi,0] \times [-5,0]) \quad \text{and}$$
$$(\theta_1, \rho_1) \sim \frac{9}{10}\text{Unif}([0,\pi] \times [0,5]) + \frac{1}{10}\text{Unif}([-\pi,0] \times [-5,0]),$$

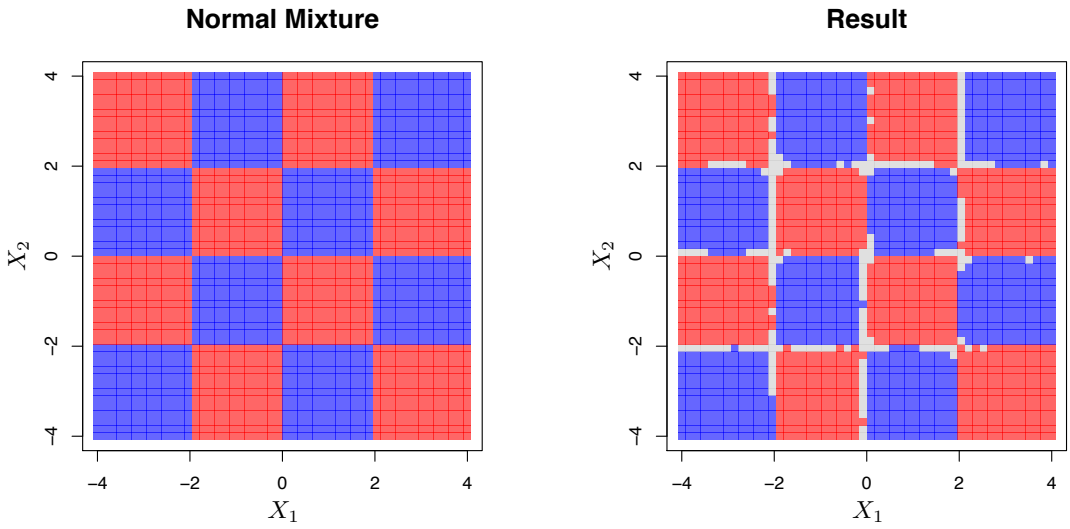

Figure 3.5: Significant local regions for the normal mixture example. The left is the underlying true model and the right is the result of the local two-sample test. The difference regions are colored as follows — (a) red: $f_1(x, y) > f_0(x, y)$, (b) blue: $f_1(x, y) < f_0(x, y)$ and (c) gray: insignificant.

and generate corresponding edge images. As a result, there are two sets of the edge images supported on $\mathbb{R}^{256}$. Using these image samples, we implemented Algorithm 2 to detect local significant points. The local tests were performed at fixed images whose parameters are defined on a uniform grid of 200×200 points over $(\theta, \rho) \in [-\pi, \pi] \times [-5, 5]$. For visualization purpose, we projected the testing points into the two-dimensional diffusion space (see Appendix B.2 for details) and the final result is provided in Figure 3.6.

For both examples, the kNN local regression test performs reasonably well and detects most of the local differences between two distributions.

3.6 Application to Astronomy Data

Continuing our discussion from Section 6.2, we apply our two-sample framework to galaxies in the COSMOS, EGS, GOODS-North and UDS fields observed by the Hubble Space Telescope (HST) as part of the CANDELS program.* For the analysis, we compute seven morphological statistics that summarize galaxy images nonparametrically: *M, I, D* (Freeman et al., 2013), *Gini*, M_{20} (Lotz et al., 2004), C and A (Conselice, 2003). Each statistic (see the references for details) explains particular aspects of galaxy morphology. In brief, the *M, I, D* statistics capture galaxies with disturbed morphologies, *Gini* and M_{20} describe the variance of a galaxy's stellar light distribution, and the C and A statistics measure the concentration of light and

*http://candels.ucolick.org

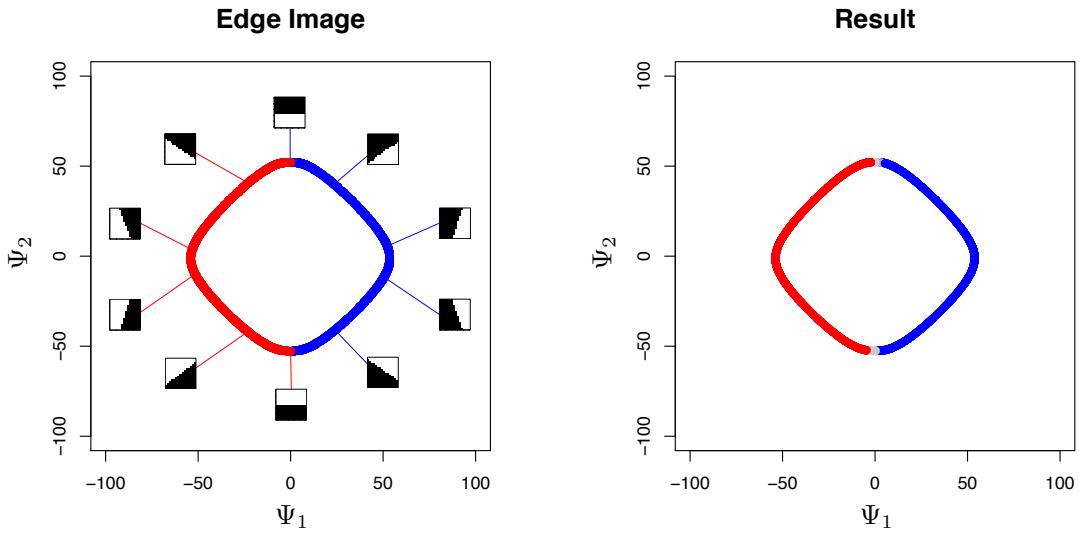

Figure 3.6: Significant local regions for the manifold data example. The left is the underlying true model and the right is the result of the local two-sample test. The difference regions are colored as follows — (a) red: $f_1(x_1, \ldots, x_{256}) > f_0(x_1, \ldots, x_{256})$, (b) blue: $f_1(x_1, \ldots, x_{256}) < f_0(x_1, \ldots, x_{256})$ and (c) gray: insignificant. Here Ψ_1 and Ψ_2 denote the the first two coordinates of the diffusion map.

asymmetry of a galaxy, respectively. We restrict our study to relatively nearby galaxy observations that have a redshift (proxy for distance) estimate between $0.56 < z < 1.12$. The final data set consists of 2736 so-called i-band-selected galaxy observations. For each galaxy, we have seven morphological image statistics along with an estimate of star-formation rate (SFR).

Galaxy morphology is closely related to other physical properties such as star formation rate, mass and metallicity (see, e.g., Snyder et al., 2015). The aim of this study is to demonstrate that our local two-sample framework can be valuable in detecting and quantifying dependencies between variables of moderate or high dimension without resorting to low-dimensional projections of summary statistics. In particular, we demonstrate that local two-sample tests can identify galaxies that lie in regions of the feature space where the estimated proportion of a particular defined class of objects (such as star-forming galaxies) differs significantly from the global proportion. Hence, we start by defining two galaxy classes based on the SFR: we say that a galaxy belongs to the high-SFR group if its SFR is higher than the upper 25% quantile of the SFR distribution ($\log_{10}(\text{SFR}) > 1.201$), and that it belongs to the low-SFR group if its SFR is lower than the lower 25% quantile of the SFR distribution ($\log_{10}(\text{SFR}) < -0.915$). We further randomly divide the data into a training set ($n = 684$) and a test set ($n = 684$). We use the training data to construct the local test statistic in (3.3), and we perform the local-two sample tests at the points in the test set (that is, these are the evaluation points in Algorithm 2). Note that this particular application is especially challenging because

the seven morphological statistics have very different properties, and some of the statistics (M and I) are essentially of mixed discrete and continuous type with heavy outliers; hence, any metric-based estimator is bound to perform poorly even after normalizing the variables. Our regression test, however, can by-pass this problem by leveraging the random forest algorithm. Another advantage of using random forests is that the algorithm returns variable importance measures that can help us identify *which* morphology statistics are the most important in distinguishing the two populations (Figure 3.7).

3.6.1 Analysis and Result

According to our global two-sample test ($\widehat{\mathcal{T}}_{RF} = .188$, $p < .001$), there is a significant difference between the low-SFR and the high-SFR populations in terms of galaxy morphology. We follow up on this result by implementing the local two-sample testing framework according to Algorithm 2 with FWER control at $\alpha = 0.05$ by the Hochberg step up procedure. To visualize locally significant points from the local test, we use diffusion maps with local scaling (Zelnik-Manor and Perona, 2005). For more information on our particular application of diffusion maps, see Appendix B.2. The main result of the local significance test is displayed in Figure 3.1. As we can see, the high-SFR and low-SFR dominated regions (that is, the regions where $f_{\text{LowSFR}} < f_{\text{HighSFR}}$ and $f_{\text{LowSFR}} > f_{\text{HighSFR}}$, respectively) are fairly well-separated in morphology space. Figure 3.1 also shows some examples of galaxy images at significant test points. By inspecting such images, we note that the "red" galaxies in the low-SFR dominated regions of the seven-dimensional space tend to be more concentrated and less disturbed than their "blue" counterparts in the high-SFR dominated regions — this result is consistent with previous astronomical studies about irregular galaxies displaying merger activities and high star-formation rates. Our test result is further supported by the variable importance measures in Figure 3.7: the two most important morphology statistics in distinguishing between high-SFR and low-SFR galaxies are the *Gini* (Lotz et al., 2004) and I (Freeman et al., 2013) morphology statistics. Indeed, by definition, the *Gini* statistic describes the variance of a galaxy's stellar light distribution, and the I statistic captures galaxies with disturbed morphologies.

3.7 Conclusions

In this work, we presented a new framework for both global and local two-sample testing via regression. Depending on the chosen regression model, our framework can efficiently deal with different types of variables and different structures in the data; thereby, providing tests with competitive power against many practical alternatives. Compared to other recent approaches in the two-sample literature (such as classification tests), our framework has the key advantage of being able to detect locally significant regions in multivariate spaces. Throughout this work, we studied theoretical properties of the regression tests by building on

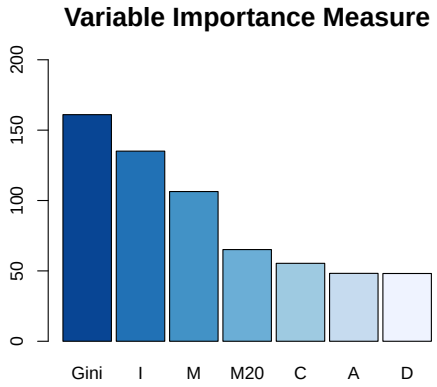

Figure 3.7: Variable importance measures from random forest regression, as measured by the Mean Decrease Gini (MDG) metric when splitting the data along the indicated variables. For the morphology-SFR study, the *Gini* and I morphology statistics are the two most important features in distinguishing between high-star-forming and the low-star-forming galaxy populations.

existing regression results. We established a connection between the power of the global and local tests to the MISE and MSE of the corresponding regression estimators, and we demonstrated practical usefulness of our methods via simulations.

By taking advantage of permutation tests under the global null hypothesis, the proposed local testing framework ensures that the type I error rate is less than or equal to the significance level. When the local null hypothesis $H_0(x) : m(x) = \pi$ is of interest, on the other hand, there is no such guarantee. In this case, it would be necessary to use an asymptotic framework and investigate the limiting behavior of a local test statistic. This topic is reserved for future work. Another direction for future work is to study the optimality of global regression tests. Contrary to the local regression test, a regression estimator with the optimal estimation error rate may not necessarily return minimax optimal global regression test. We hope that future studies will establish a lower bound and matching upper bound for the global regression test.

Chapter 4

Robust Multivariate Nonparametric Tests via Projection-Averaging

This chapter is adapted from my joint work with Sivaraman Balakrishnan and Larry Wasserman. This work is accepted to *the Annals of Statistics* for publication.

4.1 Introduction

Let X and Y be random vectors defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with distributions P_X and P_Y, respectively. Given two mutually independent samples $\mathcal{X}_m = \{X_1, \ldots, X_m\}$ and $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ from P_X and P_Y, we want to test

$$H_0 : P_X = P_Y \quad \text{versus} \quad H_1 : P_X \neq P_Y. \tag{4.1}$$

This fundamental problem has received considerable attention in statistics with a wide range of applications (see e.g. Thas, 2010, for a review). A common statistic for the univariate two-sample testing is the Cramér–von Mises (CvM) statistic (Anderson, 1962):

$$\frac{mn}{m+n} \int_{\infty}^{\infty} \big(\widehat{F}_X(t) - \widehat{F}_Y(t)\big)^2 d\widehat{H}(t),$$

where $\widehat{F}_X(t)$ and $\widehat{F}_Y(t)$ are the empirical distribution functions of $\mathcal{X}_m$ and $\mathcal{Y}_n$, respectively, and $(m+n)\widehat{H}(t) = m\widehat{F}_X(t) + n\widehat{F}_Y(t)$. Another approach is based on the energy statistic, which is an estimate of the squared

energy distance (Székely and Rizzo, 2013):

$$E^2 = 2\mathbb{E}[|X_1 - Y_1|] - \mathbb{E}[|X_1 - X_2|] - \mathbb{E}[|Y_1 - Y_2|],$$

where $|x|$ is the absolute value of $x \in \mathbb{R}$. The energy distance is well-defined assuming a finite first moment and it can be written in a form that is similar to Cramér's distance (Cramér, 1928), namely,

$$E^2 = 2\int_{-\infty}^{\infty} \big(F_X(t) - F_Y(t)\big)^2 dt,$$

where $F_X(t)$ and $F_Y(t)$ are the distribution functions of X and Y, respectively.

The CvM-statistic has several advantages over the energy statistic for univariate two-sample testing. For instance, the CvM-statistic is distribution-free under H_0 (Anderson, 1962) and its population counterpart is well-defined without any moment assumptions. It also has an intuitive probabilistic interpretation in terms of probabilities of concordance and discordance of four independent random variables (Baringhaus and Henze, 2017). Nevertheless, the CvM-statistic has rarely been studied for multivariate testing. A primary reason is that the CvM-statistic is essentially rank-based, which leads to a challenge to generalize it in a multivariate space. In contrast, the energy statistic can be easily applied in arbitrary dimensions as in Baringhaus and Franz (2004) and Székely and Rizzo (2004). Specifically, they defined the squared multivariate energy distance by

$$E_d^2(P_X, P_Y) = 2\mathbb{E}[\|X_1 - Y_1\|] - \mathbb{E}[\|X_1 - X_2\|] - \mathbb{E}[\|Y_1 - Y_2\|], \tag{4.2}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. The multivariate energy distance maintains the characteristic property that it is always non-negative and equal to zero if and only if $P_X = P_Y$. It can also be viewed as the average of univariate Cramér's distances of projected random variables (Baringhaus and Franz, 2004):

$$E_d^2(P_X, P_Y) = \frac{\sqrt{\pi}(d-1)\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d}{2})} \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \big(F_{\beta^\top X}(t) - F_{\beta^\top Y}(t)\big)^2 \, dt d\lambda(\beta), \tag{4.3}$$

where λ represents the uniform probability measure on the d-dimensional unit sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$, $\Gamma(\cdot)$ is the gamma function and the symbol $\top$ stands for the transpose operation.

Although the multivariate energy distance can be easily estimated in any dimension, it still requires the finite moment assumption as in the univariate case. When the underlying distributions violate this moment condition with potential outliers, the energy statistic becomes extremely unstable and the resulting test might perform poorly. Given that outlying observations arise frequently in practice with high-dimensional data, there is a need to develop a robust counterpart of the energy distance. The primary goal of this work is to introduce a robust, tuning parameter free, two-sample testing procedure that is easily applicable

in arbitrary dimensions and consistent against all fixed alternatives. Specifically, we modify the univariate CvM-statistic to generalize it to an arbitrary dimension by averaging over all one-dimensional projections. In detail, the proposed test statistic is an unbiased estimate of the squared multivariate CvM-distance defined as follows:

$$W_d^2(P_X, P_Y) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(F_{\beta^\top X}(t) - F_{\beta^\top Y}(t)\right)^2 dH_\beta(t) d\lambda(\beta), \tag{4.4}$$

where $H_\beta(t) = \vartheta_X F_{\beta^\top X}(t) + \vartheta_Y F_{\beta^\top Y}(t)$ and ϑ_X is a fixed value in $(0, 1)$ and $\vartheta_Y = 1 - \vartheta_X$. For simplicity and when there is no ambiguity, we may omit the dependency on P_X, P_Y and write $W_d(P_X, P_Y)$ as W_d.

Throughout this chapter, we refer to the process of averaging over all projections as *projection-averaging*.

4.1.1 Summary of our results

The proposed multivariate CvM-distance shares some appealing properties of the energy distance while being robust to heavy-tailed distributions or outliers. For example, W_d is invariant to orthogonal transformations and satisfies the characteristic property (Lemma 4.0.1), meaning that W_d is nonnegative and equal to zero if and only if $P_X = P_Y$. More importantly, it is straightforward to estimate W_d without using any tuning parameters (Theorem 4.1). Based on an unbiased estimate of W_d^2, we apply the permutation test procedure to determine a critical value of the test statistic. Although the permutation approach has been standard in practical implementations of two-sample testing, its theoretical properties have been less explored beyond simple cases (e.g. Pesarin, 2001). Indeed, previous studies usually consider asymptotic tests in their theory section whereas their actual tests are calibrated via permutations. We bridge the gap between theory and practice by presenting both theoretical and empirical results on the permutation test under various scenarios. Our main results regarding the CvM-distance are summarized as follows:

- **Closed-form expression** (Section 4.2): Building on Escanciano (2006) and Zhu et al. (2017), we show that the test statistic has a simple closed-form expression.
- **Asymptotic power** (Section 4.2): We prove that the permutation test based on the proposed statistic has the same asymptotic power as the oracle and asymptotic tests that assume knowledge of the underlying distributions (Section 4.2.2) against fixed and contiguous alternatives.
- **Robustness** (Section 4.3): We show that the permutation test based on the proposed statistic maintains good power in the contamination model, while the energy test becomes completely powerless in this setting.
- **Minimax optimality** (Section 4.4): We analyze the finite-sample power of the proposed permutation test and prove its minimax rate optimality against a class of alternatives that differ from the null in terms of the CvM-distance. We also show that the energy test is not optimal in our context.

- **HDLSS behavior** (Section 4.5): We consider a *high-dimension, low-sample size* (HDLSS) regime where the dimension tends to infinity while the sample size is fixed. Under this regime, we identify sufficient conditions under which the power of the proposed test converges to one. In addition, we show that the proposed test has comparable power to the high-dimensional mean tests introduced by Chen and Qin (2010) and Chakraborty and Chaudhuri (2017) under certain location models.
- **Angular distance** (Section 4.6): We introduce the angular distance between two vectors and use this to show that the multivariate CvM-distance is a special case of the generalized energy distance (Sejdinovic et al., 2013). Furthermore, the CvM-distance is the maximum mean discrepancy (Gretton et al., 2012) associated with the angular distance.

Beyond the CvM-statistic, the projection-averaging technique can be widely applicable to other nonparametric statistics. In the second part of this study, we revisit some famous univariate sign- or rank-based statistics and propose their multivariate counterparts via projection-averaging. Although there has been much effort to extend univariate sign- or rank-based statistics in a multivariate space (see e.g. Hettmansperger et al., 1998; Oja and Randles, 2004; Liu, 2006; Oja, 2010), they are either computationally expensive to implement or less intuitive to understand. Our projection-averaging approach addresses these issues by providing a tractable calculation form of statistics and by having a direct interpretation in terms of projections. In Section 4.7 and also Appendix C.5.8, we demonstrate the generality of the projection-averaging approach by presenting multivariate extensions of several existing univariate statistics.

4.1.2 Literature review

There are a number of multivariate two-sample testing procedures available in the literature. We list some fundamental methods and recent developments. Anderson et al. (1994) proposed the two-sample statistic based on the integrated square distance between two kernel density estimates. The energy statistic was introduced by Baringhaus and Franz (2004) and Székely and Rizzo (2004) independently. Biswas and Ghosh (2014) modified the energy statistic to improve the performance of the previous test for the high-dimensional location-scale and scale problems. Gretton et al. (2012) introduced a class of distances between two probability distributions, called the maximum mean discrepancy (MMD), based on a reproducing kernel Hilbert approach. Sejdinovic et al. (2013) showed that the energy distance is a special case of the MMD associated with the kernel induced by the Euclidean distance. Recently, Pan et al. (2018) proposed a new metric, named the ball divergence, between two probability distributions and connected it to the MMD. A further review of kernel-based two-sample tests can be found in Harchaoui et al. (2013).

Another line of work is based on graph constructions. Schilling (1986) and Henze (1988) introduced a multivariate two-sample test based on the k nearest neighbor (NN) graph. Mondal et al. (2015) pointed out that the previous NN test may suffer from low power for the high-dimensional location-scale problem and

provided an alternative that addresses this limitation. Another variant of the NN test, which is tailored to imbalanced samples, can be found in Chen et al. (2013). Friedman and Rafsky (1979) considered minimum spanning tree (MST) to present a generalization of the univariate run test in Wald and Wolfowitz (1940). The MST test proposed by Friedman and Rafsky (1979) has recently been modified by Chen and Friedman (2017) and Chen et al. (2018) to improve power under scale alternatives and imbalanced samples, respectively. Rosenbaum (2005) proposed a distribution-free test in finite samples based on cross-matches. More recently, Biswas et al. (2014) introduced another distribution-free test based on the shortest Hamiltonian path. A general theoretical framework for graph-based tests has been established by Bhattacharya (2018, 2019). Other recent developments include Liu and Modarres (2011), Kanamori et al. (2012), Bera et al. (2013), Lopez-Paz and Oquab (2016), Zhou et al. (2017), Mukhopadhyay and Wang (2018), among others.

The projection-averaging approach to CvM-type statistics can be found in other statistical problems. For example, Zhu et al. (1997) and Cui (2002) considered the CvM-statistic using projection-averaging to investigate one-sample goodness-of-fit tests for multivariate distributions. Escanciano (2006) proposed the CvM-based goodness-of-fit test for parametric regression models. To the best of our knowledge, however, this is the first study that investigates the CvM-statistic for the multivariate two-sample problem via projection-averaging.

Our technique to obtain a closed-form expression for projection-averaging statistics is based on Escanciano (2006). The same principle has been exploited by Zhu et al. (2017) in the context of testing for multivariate independence. We further extend the result of Escanciano (2006) to more general cases and provide an alternative proof using orthant probabilities for normal distributions.

Outline The rest of this chapter is organized as follows. In Section 4.2, we introduce our test statistic and the permutation test procedure. We then study their limiting behaviors under the conventional fixed dimension asymptotic framework. In Section 4.3, we compare the power of the CvM test with that of the energy test and highlight the robustness of the CvM test. Section 4.4 establishes minimax rate optimality of the proposed test against a certain class of alternatives associated with the CvM-distance. In Section 4.5, we study the asymptotic power of the CvM test in the HDLSS setting. We introduce the angular distance between two vectors in Section 4.6 to show that the CvM-distance is the generalized energy distance built on the introduced metric. In Section 4.7, the projection-averaging technique is applied to other sign- or rank-based statistics and this allows us to provide new multivariate extensions. Simulation results are reported in Section 8.9 to demonstrate the competitive power performance of the proposed approach with finite sample size. All proofs of the main results are deferred to the supplementary material.

Notation For two non-zero vectors $U_1, U_2 \in \mathbb{R}^d$, we denote the angle between U_1 and U_2 by $\mathsf{Ang}(U_1, U_2) = \arccos\{U_1^\top U_2/(\|U_1\|\|U_2\|)\}$. For $1 \le q \le p$, we let $(p)_q = p(p-1)\cdots(p-q+1)$. Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be the

probability measures under H_0 and H_1, respectively. Similarly $\mathbb{E}_0$ and $\mathbb{E}_1$ stand for the expectations with respect to $\mathbb{P}_0$ and $\mathbb{P}_1$. For any two real sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n \asymp b_n$ if there exist constants $C, C' > 0$ such that $C < |a_n/b_n| < C'$ for each n. We write $a_n = O(b_n)$ if there exists $C > 0$ such that $|a_n| \leq C|b_n|$ for each n. We also write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. For a sequence of random variables X_n, we use the notation $X_n = O_{\mathbb{P}}(a_n)$ when X_n is bounded in probability (tight). The acronym *i.i.d.* stands for independent and identically distributed and we use the symbol $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ to represent that $X_1, \ldots, X_n$ are *i.i.d.* samples from distribution P. We denote the $d \times d$ identity matrix by I_d. The symbol $\mathbb{1}(\cdot)$ is used for indicator functions. We write summation over the set of all k-tuples drawn without replacement from $\{1, \ldots, n\}$ by $\sum_{i_1,\ldots,i_k=1}^{n,\neq}$. Throughout this chapter, we assume that all vectors are column vectors and $m, n \geq 2$.

4.2 Projection Averaging-Type Cramér–von Mises Statistics

In this section, we start with the basic properties of the CvM-distance. We then introduce our test statistic and study its limiting behavior. We end this section with a description of the permutation test and its large sample properties. Throughout this section, we consider the conventional asymptotic regime where the dimension is fixed and

$$\frac{m}{m+n} \to \vartheta_X \in (0,1) \text{ and } \frac{n}{m+n} \to \vartheta_Y \in (0,1) \text{ as } N = m+n \to \infty. \tag{4.5}$$

Let us first establish the characteristic property of the CvM-distance.

Lemma 4.0.1 (Characteristic property)**.** *W_d is nonnegative and has the characteristic property:*

$$W_d(P_X, P_Y) = 0 \quad \textit{if and only if} \quad P_X = P_Y.$$

Note that W_d involves integration over the unit sphere. One way to approximate this integral is to consider a subset of $\mathbb{S}^{d-1}$, namely $\{\beta_1, \ldots, \beta_k\}$, and then to take the sample mean over k different univariate CvM-statistics (see e.g. Zhu et al., 1997). However, this approach has an unpleasant trade-off between accuracy and computational time depending on the choice of k. The problem becomes even worse in high dimensions where one may need exponentially many projections to achieve a certain accuracy. Our approach, on the other hand, does not suffer from this computational issue by explicitly calculating the integral over $\mathbb{S}^{d-1}$. The explicit form of the integration is mainly due to Escanciano (2006) who provided the following lemma:

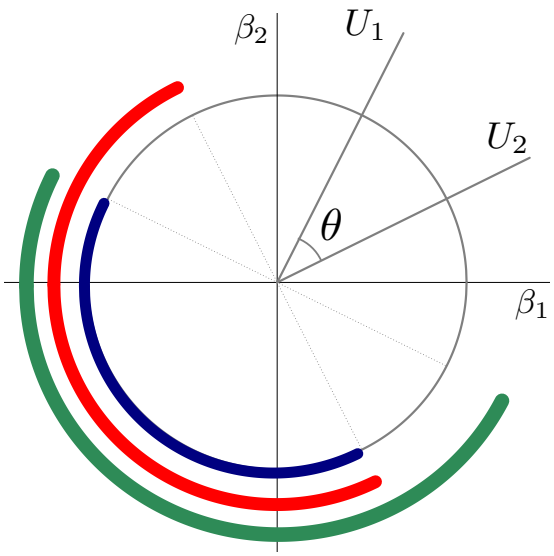

Figure 4.1: Visual proof of Lemma 4.0.2. The blue curve represents the set of $(\beta_1, \beta_2) \in \mathbb{R}^2$ that satisfies $\mathbb{1}(\beta^\top U_1 \le 0)\mathbb{1}(\beta^\top U_2 \le 0)$ and θ represents the angle between U_1 and U_2.

Lemma 4.0.2. *(Escanciano, 2006) For any non-zero vectors $U_1, U_2 \in \mathbb{R}^d$,*

$$\int_{\mathbb{S}^{d-1}} \mathbb{1}(\beta^\top U_1 \le 0)\mathbb{1}(\beta^\top U_2 \le 0) d\lambda(\beta) = \frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}\left(U_1, U_2\right).$$

Remark 4.1. Escanciano (2006) proved Lemma 4.0.2 using the volume of a spherical wedge (see Figure 4.1). In the supplementary material (Appendix C.4.2), we provide an alternative proof of this result based on orthant probabilities for normal distributions. We also extend this result to integration involving strictly more than two indicator functions in the supplementary material (Lemma C.1.8 and Lemma C.1.30). This extension allows us to generalize the univariate τ^* (Bergsma and Dassios, 2014) in Theorem 4.12 and potentially many other univariate statistics (see Appendix C.5.8).

Based on Lemma 4.0.2, we give another representation of W_d^2 in terms of the expected angle involving three independent random vectors. Here and hereafter, we assume that

$\beta^\top X$ and $\beta^\top Y$ have continuous distribution functions for λ-almost all $\beta \in \mathbb{S}^{d-1}$.

This continuity assumption greatly simplifies the alternative expression for W_d^2 and avoids the possibility that $\mathsf{Ang}(\cdot,\cdot)$ is not well-defined when one of the inputs is a zero vector. This issue may be handled by defining $\mathsf{Ang}(\cdot,\cdot)$ differently for those exceptional cases, but we do not pursue this direction here.

Theorem 4.1 (Closed-form expression)**.** *Suppose that* $X_1, X_2 \overset{i.i.d.}{\sim} P_X$ *and, independently,* $Y_1, Y_2 \overset{i.i.d.}{\sim} P_Y$*. Then the squared multivariate CvM-distance can be written as*

$$W_d^2(P_X, P_Y) \;=\; \frac{1}{3} - \frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}\left(X_1 - Y_1, X_2 - Y_1\right)\right] - \frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}\left(Y_1 - X_1, Y_2 - X_1\right)\right].$$

The above result highlights that $W_d(P_X, P_Y)$ is invariant to the choice of ϑ_X and ϑ_Y under the continuity assumption. In the next subsection, we introduce the test statistic and study its limiting behavior.

4.2.1 Test Statistic and Limiting Distributions

Theorem 4.1 leads to a natural empirical estimate of W_d^2 based on a U-statistic. Consider the kernel of order two:

$$h_{\mathrm{CvM}}(x_1, x_2; y_1, y_2) \;=\; \frac{1}{3} - \frac{1}{2\pi}\mathsf{Ang}\left(x_1 - y_1, x_2 - y_1\right) - \frac{1}{2\pi}\mathsf{Ang}\left(y_1 - x_1, y_2 - x_1\right). \tag{4.6}$$

We denote its symmetrized version, which is invariant to the order of the first two arguments as well as the last two arguments, by

$$\widetilde{h}_{\mathrm{CvM}}(x_1, x_2; y_1, y_2) \;=\; \frac{1}{2}h_{\mathrm{CvM}}(x_1, x_2; y_1, y_2) + \frac{1}{2}h_{\mathrm{CvM}}(x_2, x_1; y_2, y_1).$$

Then our test statistic is defined as follows:

$$U_{\mathrm{CvM}} = \binom{m}{2}^{-1}\binom{n}{2}^{-1} \sum_{1 \le i_1 < i_2 \le m} \sum_{1 \le j_1 < j_2 \le n} \widetilde{h}_{\mathrm{CvM}}(X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2}). \tag{4.7}$$

Leveraging the basic theory of U-statistics (e.g. Lee, 1990), it is clear that U_{CvM} is an unbiased estimator of W_d^2. Additionally, U_{CvM} is a degenerate U-statistic under the null hypothesis as we prove in the supplementary material (Appendix C.4.5). Hence we can apply the asymptotic theory for a degenerate two-sample U-statistic (Chapter 3 of Bhat, 1995) to obtain the following result.

Theorem 4.2 (Asymptotic null distribution of U_{CvM})**.** *For each* $k = 1, 2, \ldots,$ *let* λ_k *be the eigenvalue with the corresponding eigenfunction* ϕ_k *satisfying the integral equation*

$$\mathbb{E}\Big[\mathbb{E}\Big\{\widetilde{h}_{\mathrm{CvM}}(x_1, X_2; Y_1, Y_2)\Big|X_2\Big\}\phi_k(X_2)\Big] = \lambda_k\phi_k(x_1). \tag{4.8}$$

Then U_{CvM} *has the limiting null distribution under the limiting regime (4.5) given by*

$$NU_{\mathrm{CvM}} \xrightarrow{d} \vartheta_X^{-1}\vartheta_Y^{-1}\sum_{k=1}^{\infty}\lambda_k(\xi_k^2 - 1),$$

where $\xi_k \overset{i.i.d.}{\sim} N(0,1)$ *and* $\xrightarrow{d}$ *stands for convergence in distribution.*

Under a fixed alternative hypothesis where P_X and P_Y do not change with m and n, the proposed test statistic converges weakly to a normal distribution. We build on Hoeffding's decomposition of a two-sample U-statistic (e.g. page 40 of Lee, 1990) to prove the following result.

Theorem 4.3 (Asymptotic distribution of U_{CvM} under fixed alternatives)**.** *Let us define*

$$\sigma_{h_X}^2 = \mathrm{Var}\Big[\mathbb{E}\Big\{\widetilde{h}_{\mathrm{CvM}}(X_1, X_2; Y_1, Y_2)\big|X_1\Big\}\Big] \quad \textit{and}$$
$$\sigma_{h_Y}^2 = \mathrm{Var}\Big[\mathbb{E}\Big\{\widetilde{h}_{\mathrm{CvM}}(X_1, X_2; Y_1, Y_2)\big|Y_1\Big\}\Big],$$

where Var$(\cdot)$ *is the variance operator. Then under the limiting regime (4.5) and fixed alternative* $P_X \neq P_Y$*, we have*

$$\sqrt{N}(U_{\mathrm{CvM}} - W_d^2) \xrightarrow{d} N\left(0, 4\vartheta_X^{-1}\sigma_{h_X}^2 + 4\vartheta_Y^{-1}\sigma_{h_Y}^2\right).$$

From the previous two theorems, it is clear to see that NU_{CvM} is stochastically bounded under the null hypothesis whereas it diverges to infinity under fixed alternatives. Thus one can expect that any reasonable test based on the proposed test statistic is consistent (meaning that the power converges to one as $N \to \infty$) against all fixed alternatives. In fact, the problem of distinguishing two fixed distributions is too easy in large sample situations and many of nonparametric tests are known to be consistent in this asymptotic regime. We therefore turn now to a more challenging scenario where a distance between P_X and P_Y diminishes as the sample size increases. To this end, we make a standard assumption that the underlying distributions belong to quadratic mean differentiable (QMD) families (e.g. Bhattacharya, 2019).

Definition 4.1. *(Quadratic Mean Differentiable Families, page 484 of Lehmann and Romano, 2006) Let* $\{P_\theta, \theta \in \Omega\}$ *be a family of probability distributions on* $(\mathbb{R}^d, \mathcal{B})$ *where* $\mathcal{B}$ *is the Borel* σ*-field associated with* $\mathbb{R}^d$ *and* Ω *is an open subset of* $\mathbb{R}^p$*. Assume each* P_θ *is absolutely continuous with respect to Lebesgue measure and set* $p_\theta(t) = dP_\theta(t)/dt$*. The family* $\{P_\theta, \theta \in \Omega\}$ *is quadratic mean differentiable at* θ_0 *if there exists a vector of real-valued functions* $\eta(\cdot, \theta_0) = (\eta_1(\cdot, \theta_0), \ldots, \eta_p(\cdot, \theta_0))^\top$ *such that*

$$\int_{\mathbb{R}^d} \left[\sqrt{p_{\theta_0+b}(t)} - \sqrt{p_{\theta_0}(t)} - b^\top \eta(t, \theta_0)\right]^2 dt = o(\|b\|^2) \quad \textit{as } \|b\| \to 0.$$

The QMD families include a broad class of parametric distributions such as exponential families in natural form. By focusing on the QMD families, we are particularly interested in asymptotically non-degenerate situations where the limiting sum of the type I and type II errors of the optimal test is non-trivial, i.e. bounded by the nominal level α and one. It has been shown that when P_{θ_0} and P_{θ_N} belong to the QMD families, this

non-degenerate situation occurs when $\|\theta_0 - \theta_N\| \asymp N^{-1/2}$ (Chapter 13.1 of Lehmann and Romano, 2006). Hence we consider a sequence of contiguous alternatives where $\theta_N = \theta_0 + bN^{-1/2}$ for some $b \in \mathbb{R}^p$ and establish the asymptotic behavior of U_{CvM} under the given scenario. Our result builds on the prior work by Chikkagoudar and Bhat (2014) and extends it to multivariate cases.

Theorem 4.4 (Asymptotic distribution of U_{CvM} under contiguous alternatives)**.** *Assume $\{P_\theta, \theta \in \Omega\}$ is quadratic mean differentiable at θ_0 with derivative $\eta(\cdot, \theta_0)$ and Ω is an open subset of $\mathbb{R}^p$. Define the Fisher Information matrix to be the matrix $I(\theta)$ with (i, j) entry*

$$I_{i,j}(\theta) = 4 \int_{\mathbb{R}^d} \eta_i(t, \theta)\eta_j(t, \theta)dt,$$

and assume that $I(\theta_0)$ is nonsingular. Suppose we observe $\mathcal{X}_m \overset{i.i.d.}{\sim} P_{\theta_0}$ and $\mathcal{Y}_n \overset{i.i.d.}{\sim} P_{\theta_0 + bN^{-1/2}}$ for $b \in \mathbb{R}^p$. Then under the limiting regime (4.5),

$$NU_{\text{CvM}} \xrightarrow{d} \vartheta_X^{-1}\vartheta_Y^{-1} \sum_{k=1}^{\infty} \lambda_k\{(\xi_k + \vartheta_X^{1/2}a_k)^2 - 1\},$$

where $a_k = \int_{\mathbb{R}^d} 2\{b^\top \eta(x, \theta_0)\}p_{\theta_0}^{-1/2}(x)\phi_k(x)dP_{\theta_0}(x)$.

The above theorem implies that if there exists $k \geq 1$ such that $a_k \neq 0$ and $\lambda_k > 0$, a test based on U_{CvM} can have asymptotic power greater than α (see, page 615 of Lehmann and Romano, 2006). This is in contrast to the NN test which has a slower consistency rate given by $N^{-1/4}$ when $d \leq 8$ under some regularity conditions (see, Bhattacharya, 2018). In the supplementary material, we consider low-dimensional Gaussian location models and illustrate that the proposed CvM test dominates the NN test via simulations. In fact the former tends to have very close power to Hotelling's T^2 test, which is known to be optimal under the low-dimensional Gaussian scenarios (Anderson, 2003).

4.2.2 Critical Value and Permutation Test

So far we have investigated the limiting behaviors of the test statistic under the null and (fixed and contiguous) alternative hypotheses. It is important to note that the performance of a test depends not only on its test statistic but also crucially on its critical value. A common approach to determining the critical value is based on the limiting null distribution of the test statistic. Since we are dealing with a general composite null, one can define this limiting null distribution in various ways. Two natural candidates are described as follows:

- $P_{\text{CvM}}^{(\text{single})}$: the limiting distribution of NU_{CvM} based on i.i.d. samples from the single distribution P_X.

- $P_{\mathrm{CvM}}^{(\mathrm{mix})}$: the limiting distribution of NU_{CvM} based on i.i.d. samples from the mixture distribution $\vartheta_X P_X + \vartheta_Y P_Y$.

These two limiting distributions $P_{\mathrm{CvM}}^{(\mathrm{single})}$ and $P_{\mathrm{CvM}}^{(\mathrm{mix})}$ coincide when $P_X = P_Y$ but they are different in general if $P_X \neq P_Y$. By invoking Theorem 4.2, we can conclude that $P_{\mathrm{CvM}}^{(\mathrm{single})}$ and $P_{\mathrm{CvM}}^{(\mathrm{mix})}$ are Gaussian chaos distributions in the low-dimensional setting. The asymptotic tests then reject the null hypothesis when NU_{CvM} is greater than the upper $1-\alpha$ quantile of $P_{\mathrm{CvM}}^{(\mathrm{single})}$ or $P_{\mathrm{CvM}}^{(\mathrm{mix})}$, denoted by $q_{\alpha,\mathrm{CvM}}^{(\mathrm{single})}$ and $q_{\alpha,\mathrm{CvM}}^{(\mathrm{mix})}$, respectively.

Unfortunately this asymptotic approach is infeasible as the limiting distributions involve quantities that depend on the underlying distributions and that cannot be easily estimated. Even if either $P_{\mathrm{CvM}}^{(\mathrm{single})}$ or $P_{\mathrm{CvM}}^{(\mathrm{mix})}$ is known exactly, the resulting asymptotic test does not have finite-sample guarantees on the type I error control. For this reason, we advocate for using the permutation procedure that resolves the issues of the asymptotic approach. More importantly, as shown in Theorem 4.6, the power of the permutation test is asymptotically the same as that of the asymptotic tests under the conventional asymptotic regime.

Before we describe the permutation procedure, let us introduce the oracle test that serves as a benchmark for the permutation test. Let $T_{m,n}$ be a generic two-sample test statistic. Then the critical value of the oracle test based on $T_{m,n}$ can be determined as follows:

- **Oracle Test**

1. Consider new *i.i.d.* samples $\{\widetilde{Z}_1, \ldots, \widetilde{Z}_N\}$ from the mixture $\vartheta_X P_X + \vartheta_Y P_Y$.
2. Let $T_{m,n}(\widetilde{Z})$ be the test statistic of interest calculated based on $\widetilde{\mathcal{X}}_m = \{\widetilde{Z}_1, \ldots, \widetilde{Z}_m\}$ and $\widetilde{\mathcal{Y}}_n = \{\widetilde{Z}_{m+1}, \ldots, \widetilde{Z}_N\}$.
3. Given a significance level $0 < \alpha < 1$, return the critical value $c^*_{\alpha,m,n}$ defined by

$$c^*_{\alpha,m,n} := \inf\Big\{t \in \mathbb{R} : 1-\alpha \leq \mathbb{P}\Big(T_{m,n}(\widetilde{Z}) \leq t\Big)\Big\}. \tag{4.9}$$

It is worth pointing out that the oracle statistic $T_{m,n}(\widetilde{Z})$ has the same distribution as the test statistic based on the original samples under H_0, but not necessarily under H_1. Hence the oracle test based on $c^*_{\alpha,m,n}$ is exact under H_0 and can be powerful under H_1. However, $c^*_{\alpha,m,n}$ relies on the unknown mixture distribution $\vartheta_X P_X + \vartheta_Y P_Y$, which makes the oracle test impractical. In sharp contrast, the critical value of the permutation test can be obtained without knowledge of the mixture distribution as follows:

- **Permutation Test**

1. Let $\{Z_1, \ldots, Z_N\} = \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ be the pooled samples and $Z_\varpi = \{Z_{\varpi(1)}, \ldots, Z_{\varpi(N)}\}$ where $\varpi = \{\varpi(1), \ldots, \varpi(N)\}$ is a permutation of $\{1, \ldots, N\}$.

2. Let $T_{m,n}(Z_\varpi)$ be the test statistic of interest calculated based on $\mathcal{X}_m^\varpi = \{Z_{\varpi(1)}, \ldots, Z_{\varpi(m)}\}$ and $\mathcal{Y}_n^\varpi = \{Z_{\varpi(m+1)}, \ldots, Z_{\varpi(N)}\}$.

3. Given a significance level $0 < \alpha < 1$, return the critical value $c_{\alpha,m,n}$ defined by

$$c_{\alpha,m,n} := \inf\Big\{t \in \mathbb{R} : 1 - \alpha \le \frac{1}{N!}\sum_{\varpi \in \mathcal{S}_N} \mathbb{1}\Big(T_{m,n}(Z_\varpi) \le t\Big)\Big\}, \tag{4.10}$$

where $\mathcal{S}_N$ is the set of all permutations of $\{1, \ldots, N\}$.

In Theorem 4.5, we show that the difference between $c^*_{\alpha,m,n}$ and $c_{\alpha,m,n}$ for the proposed statistic is asymptotically negligible under both the null and alternative hypotheses. This connection in turn implies that the permutation critical value converges to $q^{(\text{mix})}_{\alpha,\text{CvM}}$, which is the limit of the oracle critical value by construction. Moreover, under the contiguous alternative, we also establish that $q^{(\text{single})}_{\alpha,\text{CvM}}$ is the same as $q^{(\text{mix})}_{\alpha,\text{CvM}}$. Building on this observation, we formally prove that (i) the permutation test, (ii) the oracle test and (iii) the asymptotic tests based on $P^{(\text{single})}_{\text{CvM}}$ and $P^{(\text{mix})}_{\text{CvM}}$ have the same asymptotic power against both contiguous and fixed alternatives in Theorem 4.6. In doing so, we develop a general asymptotic theory for the permutation distribution of a two-sample degenerate U-statistic under H_0. This general result is established based on Hoeffding's conditions (Hoeffding, 1952) and extended to H_1 via the coupling argument (Chung and Romano, 2013). The details can be found in Appendix C.2.

Let us denote by $c^*_{\alpha,\text{CvM}}$ and $c_{\alpha,\text{CvM}}$ the critical values of the oracle test and the permutation test based on the scaled CvM-statistic, that is NU_{CvM}, as described in the procedures (4.9) and (4.10), respectively. Then our result on the critical values is stated as follows.

Theorem 4.5 (Asymptotic behavior of the critical values)**.** *Consider the conventional limiting regime in (4.5) with the additional assumption that $m/N - \vartheta_X = O(N^{-1/2})$. Then under both the null and (fixed or contiguous) alternative hypotheses,*

$$c_{\alpha,\text{CvM}} \xrightarrow{p} q^{(\text{mix})}_{\alpha,\text{CvM}} \quad \textit{and} \quad c^*_{\alpha,\text{CvM}} \xrightarrow{p} q^{(\text{mix})}_{\alpha,\text{CvM}},$$

where $\xrightarrow{p}$ stands for convergence in probability. Moreover, under the null or contiguous alternative, we further have that $q^{(\text{mix})}_{\alpha,\text{CvM}} = q^{(\text{single})}_{\alpha,\text{CvM}}$.

Leveraging the previous result combined with Slutsky's theorem, we next prove that the asymptotic power of the oracle test, the permutation test and the asymptotic tests are identical against any fixed and contiguous alternatives. This clearly highlights an advantage of the permutation test as it is exact under H_0 and asymptotically as powerful as the oracle and asymptotic tests under H_1. More importantly the permutation test does not require any prior information on the underlying distributions.

Theorem 4.6 (Asymptotic equivalence of power)**.** *The oracle test and the permutation test control the type I error under the null hypothesis as*

$$\mathbb{P}_0\big(NU_{\mathrm{CvM}} > c^*_{\alpha,\mathrm{CvM}}\big) \le \alpha \quad \textit{and} \quad \mathbb{P}_0\big(NU_{\mathrm{CvM}} > c_{\alpha,\mathrm{CvM}}\big) \le \alpha.$$

On the other hand, under the fixed or contiguous alternative hypotheses considered in Theorem 4.3 and Theorem 4.4 with the additional assumption that $m/N - \vartheta_X = O(N^{-1/2})$*, we have*

$$\begin{aligned}
\mathbb{P}_1\big(NU_{\mathrm{CvM}} > c_{\alpha,\mathrm{CvM}}\big) &= \mathbb{P}_1\big(NU_{\mathrm{CvM}} > c^*_{\alpha,\mathrm{CvM}}\big) + o(1) \\
&= \mathbb{P}_1\big(NU_{\mathrm{CvM}} > q^{(\mathrm{single})}_{\alpha,\mathrm{CvM}}\big) + o(1) \\
&= \mathbb{P}_1\big(NU_{\mathrm{CvM}} > q^{(\mathrm{mix})}_{\alpha,\mathrm{CvM}}\big) + o(1).
\end{aligned}$$

Remark 4.2. It is worth pointing out that due to the symmetry of the kernel $\widetilde{h}_{\mathrm{CvM}}$, it is enough to consider $\binom{N}{m}$ permutations to obtain the critical value $c_{\alpha,\mathrm{CvM}}$ for the CvM test. Nevertheless, except for small sample sizes, the exact permutation procedure is too expensive to implement in practical applications. A common approach to alleviate this computational issue is to use Monte Carlo sampling of random permutations and approximate the exact permutation p-value. In more detail, note first that the permutation test function can be written as $\mathbb{1}(\widehat{p}_{\mathrm{CvM}} \le \alpha)$ where $\widehat{p}_{\mathrm{CvM}}$ is the permutation p-value given by

$$\widehat{p}_{\mathrm{CvM}} = \frac{1}{N!} \sum_{\varpi \in \mathcal{S}_N} \mathbb{1}\{U_{\mathrm{CvM}}(Z_\varpi) \ge U_{\mathrm{CvM}}\}.$$

Let $\varpi^{(1)}, \ldots, \varpi^{(B)}$ be independent and uniformly distributed on $\mathcal{S}_N$. Then the Monte Carlo version of the permutation p-value is computed by

$$\widehat{p}^{(B)}_{\mathrm{CvM}} = \frac{1}{B+1}\left[\sum_{i=1}^{B} \mathbb{1}\{U_{\mathrm{CvM}}(Z_{\varpi^{(i)}}) \ge U_{\mathrm{CvM}}\} + 1\right].$$

It is well-known that $\mathbb{1}(\widehat{p}^{(B)}_{\mathrm{CvM}} \le \alpha)$ is also a valid level α test for any finite sample size and $\widehat{p}_{\mathrm{CvM}} - \widehat{p}^{(B)}_{\mathrm{CvM}} \xrightarrow{p} 0$ as $B \to \infty$ (e.g. page 636 of Lehmann and Romano, 2006). Throughout this chapter, we also adopt this approach for our simulation studies.

4.3 Robustness

Recall that the energy distance and the CvM-distance can be represented by integrals of the L_2^2-type difference between two distribution functions. In view of this, the main difference between the energy distance and

the CvM-distance is in their weight function. More precisely, the energy distance is defined with dt, which gives a uniform weight to the whole real line. On the other hand, the CvM-distance is defined with $dH_\beta(t)$, which gives the most weight on high-density regions. As a result, the test based on the CvM-distance is more robust to extreme observations than the one based on the energy distance. It is also important to note that the CvM-distance is well-defined without any moment conditions, whereas the energy distance is only well-defined assuming a finite first moment. When the moment condition is violated or there exist extreme observations, the test based on the energy distance may perform poorly. The purpose of this section is to demonstrate this point both theoretically and empirically by using contaminated distribution models.

4.3.1 Theoretical Analysis

Suppose that we observe samples from an ϵ-contamination model:

$$\begin{aligned} X &\sim P_{X,N} := (1-\epsilon)Q_X + \epsilon G_N \quad \text{and} \\ Y &\sim P_{Y,N} := (1-\epsilon)Q_Y + \epsilon G_N, \end{aligned} \tag{4.11}$$

where G_N can change arbitrarily with N and $\epsilon \in (0,1)$. Suppose that Q_X and Q_Y are different so that a given test has high power to distinguish between Q_X and Q_Y without contaminations. Then it is natural to expect that the power of the same test would not decrease much for the contamination model when ϵ is close to zero. In other words, an ideal test would maintain robust power against any choice of G_N as long as Q_X and Q_Y are different and ϵ is small. Unfortunately, this is not the case for the energy test. As we shall see, for any arbitrary small (but fixed) ϵ, there exists a contamination G_N such that the energy test becomes asymptotically powerless under mild moment conditions for Q_X and Q_Y. On the other hand, the CvM test is uniformly powerful over any choice of G_N as sample size tends to infinity.

Remark 4.3. We mainly focus on statistical power to study robustness because one can always employ the permutation procedure to control the type I error under $H_0 : P_{X,N} = P_{Y,N}$.

Let us consider the energy statistic based on a U-statistic:

$$U_{\text{Energy}} = \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\|X_i - Y_j\| - \frac{1}{(m)_2}\sum_{i_1,i_2=1}^{m,\neq}\|X_{i_1} - X_{i_2}\| - \frac{1}{(n)_2}\sum_{j_1,j_2=1}^{n,\neq}\|Y_{j_1} - Y_{j_2}\|. \tag{4.12}$$

Then the main result of this subsection is stated as follows.

Theorem 4.7 (Robustness under contaminations)**.** *Suppose we observe samples $\mathcal{X}_m$ and $\mathcal{Y}_n$ from the contaminated model in (4.11) with an arbitrary small but fixed contamination ratio ϵ. Assume that Q_X and Q_Y are fixed but $Q_X \neq Q_Y$ while N changes. In addition, assume that Q_X and Q_Y have their finite*

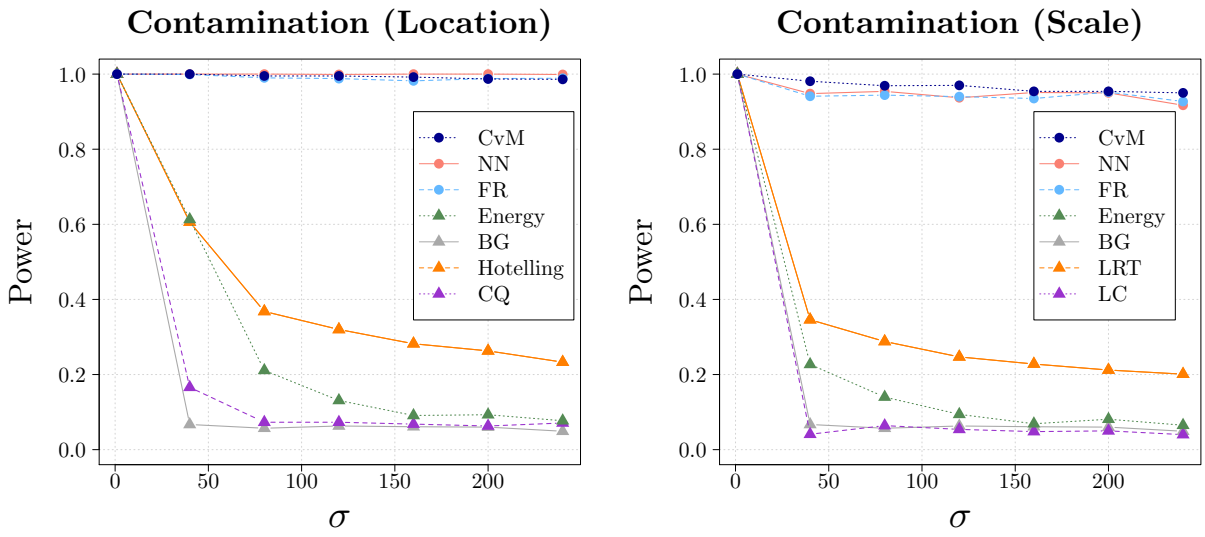

Figure 4.2: Empirical power of NN, FR, Energy, BG, Hotelling, CQ, LRT, LC and CvM tests under the contamination models with $\epsilon = 0.05$. See Example 4.1 and 4.2 for details.

second moments. Consider the tests based on U_{CvM} *and* U_{Energy} *given by*

$$\phi_{\text{CvM}} := \mathbb{1}(U_{\text{CvM}} > c_{\alpha,\text{CvM}}) \quad \textit{and} \quad \phi_{\text{Energy}} := \mathbb{1}(U_{\text{Energy}} > c_{\alpha,\text{Eng}}),$$

where $c_{\alpha,\text{CvM}}$ *and* $c_{\alpha,\text{Eng}}$ *are* α *level permutation critical values of* U_{CvM} *and* U_{Energy} *respectively. Then for any* (Q_X, Q_Y)*, there exists a certain* G_N *such that the energy test becomes asymptotically powerless under the asymptotic regime in (4.5), whereas the CvM test is asymptotically powerful uniformly over all possible* G_N*. More precisely,*

$$\lim_{m,n\to\infty} \inf_{G_N} \mathbb{E}_1 [\phi_{\text{Energy}}] \le \alpha \quad \textit{and} \quad \lim_{m,n\to\infty} \inf_{G_N} \mathbb{E}_1 [\phi_{\text{CvM}}] = 1. \tag{4.13}$$

Remark 4.4. In Theorem 4.7, we made the assumption that Q_X and Q_Y are fixed and have finite second moments. We also assumed the asymptotic regime in (4.5). These assumptions are mainly for the energy test and are not necessary for the CvM test. In fact, the same result can be derived for the CvM test given that there is a positive sequence $b_{m,n} \to \infty$ increasing arbitrary slowly with m, n such that $W_d(Q_X, Q_Y) \ge b_{m,n}(1/\sqrt{m} + 1/\sqrt{n})$ (see Theorem 5.5).

4.3.2 Empirical Analysis

To illustrate Theorem 4.7 with finite sample size, we carried out simulation studies using the contamination model in (4.11). In our simulation, we take Q_X and Q_Y to have multivariate normal distributions with

different location parameters or different scale parameters. In both examples, we take G_N to have a multivariate normal distribution given by

$$G_N := N((0, \ldots, 0)^\top, \sigma^2 I_d),$$

where σ controls the scale of the contamination G_N.

Example 4.1 (Location difference)**.** *For the location alternative, we compare two multivariate normal distributions, where the means are different but the covariance matrices are identical. Specifically, we set*

$$Q_X = N((-0.5, \ldots, -0.5)^\top, I_d), \quad \textit{and} \quad Q_Y = N((0.5, \ldots, 0.5)^\top, I_d),$$

with $\epsilon = 0.05$*. We then change* $\sigma = 1, 40, 80, 120, 160, 200$ *and* 240 *to investigate the robustness of the tests against contamination with large scale parameter* σ*.*

Example 4.2 (Scale difference)**.** *Similar to the location alternative, we again choose multivariate normal distributions which differ in their scale but not in their location parameters. In detail, we have*

$$Q_X = N((0, \ldots, 0)^\top, 0.1^2 \times I_d), \quad \textit{and} \quad Q_Y = N((0, \ldots, 0)^\top, I_d),$$

with $\epsilon = 0.05$*. Again, we change* $\sigma = 1, 40, 80, 120, 160, 200$ *and* 240 *to assess the effect of contamination with large scale parameter* σ*.*

In addition to the energy test, we further considered three nonparametric tests in our simulation studies, namely, the k-nearest neighbor test by Schilling (1986) with $k = 3$, the MST test proposed by Friedman and Rafsky (1979) and the inter-point distance test by Biswas and Ghosh (2014). For future reference, we refer to them as the NN test, the FR test and the BG test, respectively. We also added the high-dimensional mean test by Chen and Qin (2010) and Hotelling's T^2 test (e.g. page 188 of Anderson, 2003) for the location alternative and the high-dimensional covariance test by Li and Chen (2012) and the conventional likelihood ratio test (e.g. page 412 of Anderson, 2003) for the scale alternative. We refer to them as the CQ test, Hotelling's test, the LC test and the LRT test, respectively.

Experiments were run $1,000$ times to estimate the power of different tests with $m = n = 40$ and $d = 10$ at significance level $\alpha = 0.05$. The p-value of each test was computed using 500 permutations as in Remark 4.2. As can be seen from Figure 4.2, the power of the CvM test is consistently robust to the value of σ, which supports our theoretical result. The power of the energy test, on the other hand, drops down significantly as σ increases for both location and scale differences. As explained in the proof of Theorem 4.7, this poor performance was attributed to the fact that the energy statistic is very much dominated by extreme observations from G_N when σ is large. The graph-based tests, i.e. the NN and FR tests, also show a robust

power performance against the contamination models. Intuitively speaking, they perform robustly under the given scenarios as their test statistics, which count the number of edges in a graph, do not vary a lot even in the presence of outliers; but as far as we know, there is no theoretical support for this result in the current literature. Moreover these graph-based tests typically exhibit poorer consistency rates (Bhattacharya, 2018) compared to the proposed CvM test. The other four tests (Hotelling's test, the LRT test, the LC test and the CQ test) perform poorly for large σ, which may be explained similarly as to why the energy test has low power in these examples.

4.4 Minimax Optimality

Although our choice of the U-statistic was a natural one to estimate W_d^2, it remains unclear whether one can come up with a better test statistic for testing whether $H_0 : W_d = 0$ or $H_0 : W_d > 0$. One might also wonder whether there exists a testing procedure that leads to significantly higher power than the permutation test while controlling the type I error. In this section, we shall show that the answer is negative from a minimax point of view. In particular, we prove that the permutation test based on U_{CvM} is minimax rate optimal against a class of alternatives associated with the CvM-distance.

To formulate the minimax problem, let us define the set of two multivariate distributions which are at least ϵ far apart in terms of the CvM-distance, i.e.

$$\mathcal{F}(\epsilon) := \big\{(P_X, P_Y) : W_d(P_X, P_Y) \geq \epsilon\big\}.$$

For a given significance level $\alpha \in (0, 1)$, let $\mathbb{T}_{m,n}(\alpha)$ be the set of measurable functions $\phi : \{\mathcal{X}_m, \mathcal{Y}_n\} \mapsto \{0, 1\}$ such that

$$\mathbb{T}_{m,n}(\alpha) = \{\phi : \mathbb{P}_0(\phi = 1) \leq \alpha\}.$$

We then define the minimax type II error as follows:

$$1 - \beta_{m,n}(\epsilon) = \inf_{\phi \in \mathbb{T}_{m,n}(\alpha)} \sup_{P_X, P_Y \in \mathcal{F}(\epsilon)} \mathbb{P}_1(\phi = 0). \tag{4.14}$$

Our primary interest is in finding the minimax separation $\epsilon_{m,n}$ satisfying

$$\epsilon_{m,n} = \inf\big\{\epsilon : 1 - \beta_{m,n}(\epsilon) \leq \zeta\big\},$$

for some $0 < \zeta < 1 - \alpha$. We start by establishing a lower bound for the minimax separation $\epsilon_{m,n}$ based on Neyman–Pearson lemma.

Theorem 4.8 (Lower Bound). *For $0 < \zeta < 1 - \alpha$, there exists some constant $b = b(\alpha, \zeta)$ independent of the dimension such that $\epsilon_{m,n} = b(m^{-1/2} + n^{-1/2})$ and the minimax type II error is lower bounded by ζ, i.e.*

$$1 - \beta_{m,n}\left(\epsilon_{m,n}\right) \geq \zeta.$$

The above result shows that if $\epsilon_{m,n}$ is of lower order than $m^{-1/2} + n^{-1/2}$, then no test has the type II error that is uniformly smaller than the nominal level α. We now prove that this lower bound is tight by establishing a matching upper bound. In particular, the upper bound is obtained by the permutation test based on U_{CvM}, highlighting that the proposed approach is minimax rate optimal.

Theorem 4.9 (Upper Bound). *Recall the CvM test ϕ_{CvM} given in Theorem 4.7. For a sufficiently large $c > 0$, let $\epsilon^{\star}_{m,n}$ be the radius of interest defined by*

$$\epsilon^{\star}_{m,n} := c\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right). \tag{4.15}$$

Then there exists $\zeta \in (0, 1 - \alpha)$ such that the type II error of ϕ_{CvM} is uniformly bounded by ζ, i.e.

$$\sup_{P_X, P_Y \in \mathcal{F}(\epsilon^{\star}_{m,n})} \mathbb{P}_1\left(\phi_{\mathrm{CvM}} = 0\right) < \zeta.$$

Remark 4.5. We would like to emphasize that no assumption has been made in Theorem 5.5 regarding the ratio of the sample sizes. This implies that the proposed test can be consistent against general alternatives even when the two sample sizes are highly unbalanced as $m/n \to 0$ or $m/n \to \infty$.

As a straightforward consequence of Theorem 4.7, we also show that the energy test, which is our main competitor, is not minimax rate optimal in our context.

Proposition 4.1 (Non-optimality of the energy test). *Recall the energy test ϕ_{Energy} given in Theorem 4.7. Then there exists a pair of distributions that belongs to $\mathcal{F}(\epsilon^{\star}_{m,n})$ such that the energy test becomes asymptotically powerless, i.e.*

$$\lim_{m,n \to \infty} \inf_{P_X, P_Y \in \mathcal{F}(\epsilon^{\star}_{m,n})} \mathbb{P}_1(\phi_{\mathrm{Energy}} = 1) \leq \alpha.$$

In the next section we turn our attention to the asymptotic regime where the sample size is fixed and the dimension tends to infinity and study the limiting behavior of the CvM test.

4.5 High Dimension, Low Sample Size Analysis

The high dimension, low sample size (HDLSS) regime has received increasing attention in recent years and has been frequently employed to give statistical insights into high-dimensional two-sample testing (e.g. Biswas and Ghosh, 2014; Biswas et al., 2014; Mondal et al., 2015; Chakraborty and Chaudhuri, 2017). Focusing on this HDLSS regime, the goal of this section is twofold: Firstly, we provide sufficient conditions under which the proposed test is consistent in HDLSS situations (Section 4.5.1). Secondly, we show that U_{CvM} has the same asymptotic behavior as the high-dimensional mean test statistics proposed by Chen and Qin (2010) and Chakraborty and Chaudhuri (2017) under certain location models (Section 4.5.2). Along with these mean test statistics, we further establish the equivalence among U_{CvM}, the energy statistic and the MMD statistic with the Gaussian kernel. The latter connection was motivated by Ramdas et al. (2015) who showed that the energy statistic, the MMD statistic and the mean test statistic by Chen and Qin (2010) are asymptotically equivalent in different scenarios.

4.5.1 HDLSS Consistency

Let us denote $\mathbb{E}(X) = \mu_X$, $\mathbb{E}(Y) = \mu_Y$, $\mathrm{Var}(X) = \Sigma_X$ and $\mathrm{Var}(Y) = \Sigma_Y$ where Σ_X and Σ_Y are positive definite matrices. Before presenting the main results, we state the two assumptions.

(A1). $\mathrm{Var}(\|Z_1^* - Z_2^*\|^2) = O(d)$, and $\mathrm{Var}\{(Z_1^* - Z_3^*)^\top (Z_2^* - Z_3^*)\} = O(d)$ where Z_1^*, Z_2^*, Z_3^* are independent and each Z_i^* follows either P_X or P_Y.

(A2). $d^{-1}\mathrm{tr}(\Sigma_X) \to \overline{\sigma}_X^2$, $d^{-1}\mathrm{tr}(\Sigma_Y) \to \overline{\sigma}_Y^2$, $d^{-1}\|\mu_X - \mu_Y\|_2^2 \to \overline{\delta}_{XY}^2$ where $0 < \overline{\sigma}_X^2, \overline{\sigma}_Y^2 < \infty$ and $0 \leq \overline{\delta}_{XY}^2 < \infty$.

Assumption **(A1)** implies that component variables are weakly dependent. Under the distributional assumptions (including multivariate normal distributions) made in Bai and Saranadasa (1996) and Chen and Qin (2010), Assumption **(A1)** is satisfied when

$$\begin{aligned} &(\mu_X - \mu_Y)^\top (\Sigma_X + \Sigma_Y)(\mu_X - \mu_Y) = O(d) \quad \text{and} \\ &\mathrm{tr}\{(\Sigma_X + \Sigma_Y)^2\} = O(d). \end{aligned} \tag{4.16}$$

The details of this derivation can be found in Appendix C.5.1. Assumption **(A2)** is common in the HDLSS literature (e.g. Hall et al., 2005) and facilitates the analysis. Under these two conditions, the following theorem establishes the HDLSS consistency of the proposed test where we assume that the nominal level satisfies $\alpha > 1/\{(m+n)!/(m!n!)\}$ for $m \neq n$ and $\alpha > 2/\{(m+n)!/(m!n!)\}$ for $m = n$.

Theorem 4.10 (HDLSS consistency). *Suppose* **(A1)** *and* **(A2)** *hold. Assume that* $\overline{\sigma}_X^2 \neq \overline{\sigma}_Y^2$ *or* $\overline{\delta}_{XY}^2 > 0$. *Then the permutation test based on* U_{CvM} *is consistent under the HDLSS regime, that is* $\lim_{d\to\infty} \mathbb{E}_1[\phi_{\mathrm{CvM}}] = 1$.

4.5.2 HDLSS Asymptotic Equivalence of CvM-statistic and Others

Next we focus on mean difference alternatives with equal covariance matrices. There are many types of high-dimensional mean inference procedures in the literature (see Hu and Bai, 2016, for a recent review). For example, Chen and Qin (2010) suggest a test statistic based on an unbiased estimator of $\|\mu_X - \mu_Y\|^2$. Specifically, their test statistic is given by

$$U_{\mathrm{CQ}} = \frac{1}{(m)_2(n)_2} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j_1,j_2=1}^{n,\neq} (X_{i_1} - Y_{j_1})^\top (X_{i_2} - Y_{j_2}).$$

More recently, Chakraborty and Chaudhuri (2017) define a test statistic based on spatial ranks as

$$U_{\mathrm{WMW}} = \frac{1}{(m)_2(n)_2} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j_1,j_2=1}^{n,\neq} \frac{(X_{i_1} - Y_{j_1})}{\|X_{i_1} - Y_{j_1}\|}^\top \frac{(X_{i_2} - Y_{j_2})}{\|X_{i_2} - Y_{j_2}\|}.$$

They proved that U_{CQ} and U_{WMW} are asymptotically equivalent under a certain HDLSS setting. Independently, the equivalence between U_{CQ}, U_{Energy} and the MMD statistic with the Gaussian kernel was established by Ramdas et al. (2015) under different settings. Let us denote the MMD statistic with the Gaussian kernel by

$$\begin{aligned} U_{\mathrm{MMD}} = {} & \frac{1}{(m)_2} \sum_{i_1,i_2=1}^{m,\neq} \exp\Big(-\frac{1}{2\varsigma_d^2}\|X_{i_1} - X_{i_2}\|^2 \Big) \\ & + \frac{1}{(n)_2} \sum_{j_1,j_2=1}^{n,\neq} \exp\Big(-\frac{1}{2\varsigma_d^2}\|Y_{j_1} - Y_{j_2}\|^2 \Big) - \frac{2}{mn} \sum_{i=1}^{m}\sum_{j=1}^{n} \exp\Big(-\frac{1}{2\varsigma_d^2}\|X_i - Y_j\|^2 \Big), \end{aligned}$$

where ς_d^2 is the bandwidth parameter. Here we combine and further extend these results by presenting sufficient conditions under which U_{CvM}, U_{Energy}, U_{MMD}, U_{CQ} and U_{WMW} are asymptotically equivalent. To establish the result, we need two more assumptions.

(A3). $\mathrm{Var}\{(Z_1^* - Z_2^*)^\top (Z_3^* - Z_4^*)\} = O(d)$ where $Z_1^*, Z_2^*, Z_3^*, Z_4^*$ are independent and each Z_i^* follows either P_X or P_Y.

(A4). $\Sigma_X = \Sigma_Y$ and $\|\mu_X - \mu_Y\|^2 = O(\sqrt{d})$.

Assumption **(A3)** is required for studying U_{CQ} and U_{WMW}. Like Assumption **(A1)**, Assumption **(A3)** is also satisfied under condition (4.16). Notice that U_{CQ} and U_{WMW} are only sensitive to location parameters

whereas U_{CvM}, U_{Energy} and U_{MMD} are sensitive to both location and scale parameters. This suggests that the equal covariance assumption in **(A4)** is crucial for our result and cannot be dropped. The condition $\|\mu_X - \mu_Y\|^2 = O(\sqrt{d})$ is also important for our analysis and was also considered in Chakraborty and Chaudhuri (2017). Under the given assumptions, we make repeated use of Taylor expansions to establish the equivalence among the test statistics stated as follows.

Theorem 4.11 (HDLSS equivalence)**.** *Suppose* **(A1)**, **(A2)**, **(A3)** *and* **(A4)** *hold. Let ϖ be an arbitrary permutation of $\{1, \ldots, N\}$ and $\overline{\sigma}_d^2 = d^{-1}\mathrm{tr}(\Sigma_X)$. We denote by $U^{\varpi}_{\mathrm{CvM}}$, $U^{\varpi}_{\mathrm{Energy}}$, $U^{\varpi}_{\mathrm{MMD}}$, U^{ϖ}_{CQ} and $U^{\varpi}_{\mathrm{WMW}}$, the CvM, Energy, MMD, CQ, and WMW test statistics, respectively, calculated based on $\mathcal{X}_m^{\varpi} = \{Z_{\varpi(1)}, \ldots, Z_{\varpi(m)}\}$ and $\mathcal{Y}_n^{\varpi} = \{Z_{\varpi(m+1)}, \ldots, Z_{\varpi(N)}\}$. Assume that the bandwidth parameter of the Gaussian kernel satisfies $\varsigma_d^2 \asymp d$. Then under the HDLSS asymptotics, we have that*

$$\begin{aligned}
\sqrt{d}U^{\varpi}_{\mathrm{CvM}} &= \frac{1}{2\pi\sqrt{3d\overline{\sigma}_d^2}}U^{\varpi}_{\mathrm{CQ}} + O_{\mathbb{P}}(d^{-1/2}),\ U^{\varpi}_{\mathrm{Energy}} = \frac{1}{\sqrt{2d\overline{\sigma}_d}}U^{\varpi}_{\mathrm{CQ}} + O_{\mathbb{P}}(d^{-1/2}),\\
\sqrt{d}U^{\varpi}_{\mathrm{WMW}} &= \frac{1}{\sqrt{d\overline{\sigma}_d^2}}U^{\varpi}_{\mathrm{CQ}} + O_{\mathbb{P}}(d^{-1/2}),\ \sqrt{d}U^{\varpi}_{\mathrm{MMD}} = \frac{\sqrt{d}}{\varsigma_d^2}e^{-d\overline{\sigma}_d^2/\varsigma_d^2}U^{\varpi}_{\mathrm{CQ}} + O_{\mathbb{P}}(d^{-1/2}).
\end{aligned} \tag{4.17}$$

Note that the asymptotic equivalence established in (4.17) holds for any permutation. Leveraging this result, we show that the permutation critical values of the test statistics are asymptotically the same as well.

Corollary 4.11.1 (Permutation critical values)**.** *Consider the same assumptions made in Theorem 4.11. Let $c_{\alpha,\mathrm{CvM}}$, $c_{\alpha,\mathrm{Eng}}$, $c_{\alpha,\mathrm{MMD}}$, $c_{\alpha,\mathrm{CQ}}$ and $c_{\alpha,\mathrm{WMW}}$ be the $1-\alpha$ quantile of the permutation distribution of $2\pi\sqrt{3d\overline{\sigma}_d^2}U_{\mathrm{CvM}}$, $\sqrt{2}\overline{\sigma}_d U_{\mathrm{Energy}}$, $\varsigma_d^2 e^{-d\overline{\sigma}_d^2/\varsigma_d^2}U_{\mathrm{MMD}}/\sqrt{d}$, $U_{\mathrm{CQ}}/\sqrt{d}$ and $\sqrt{d\overline{\sigma}_d^2}U_{\mathrm{WMW}}$, respectively. Then*

$$\begin{aligned}
c_{\alpha,\mathrm{CvM}} &= c_{\alpha,\mathrm{Eng}} + O_{\mathbb{P}}(d^{-1/2}) = c_{\alpha,\mathrm{MMD}} + O_{\mathbb{P}}(d^{-1/2})\\
&= c_{\alpha,\mathrm{CQ}} + O_{\mathbb{P}}(d^{-1/2}) = c_{\alpha,\mathrm{WMW}} + O_{\mathbb{P}}(d^{-1/2}).
\end{aligned}$$

From the previous results, we expect that the considered permutation tests have comparable power in the limit as further illustrated by our simulation results in Section 8.9. We also refer the reader to Appendix C.5.5 where we present an explicit expression for the limiting power function of the asymptotic tests with extra assumptions. We would like to emphasize, however, that when the moment assumption is violated, the power of these tests can be entirely different. For instance, our simulation results in Section 8.9 demonstrate that the CQ, energy and MMD tests perform poorly when X and Y have Cauchy distributions with different location parameters. In contrast, the CvM and WMW tests maintain robust power against the same Cauchy alternative, which highlights a benefit of the current approach in high dimensions.

4.6 Connection to the Generalized Energy Distance and MMD

Recall that the energy distance is defined with the Euclidean distance under the finite first moment condition. By considering a semimetric space $(\mathbb{Z}, \rho)$ of negative type, Sejdinovic et al. (2013) generalized the energy distance by

$$E_\rho^2 = 2\mathbb{E}[\rho(X_1, Y_1)] - \mathbb{E}[\rho(X_1, X_2)] - \mathbb{E}[\rho(Y_1, Y_2)].$$

They further established the equivalence between the generalized energy distance and the MMD with a kernel induced by $\rho(\cdot, \cdot)$. Given a distance-induced kernel $k(\cdot, \cdot)$, the squared MMD is given by

$$\mathrm{MMD}_k^2 = \mathbb{E}[k(X_1, X_2)] + \mathbb{E}[k(Y_1, Y_2)] - 2\mathbb{E}[k(X_1, Y_1)].$$

In this section, we show that the multivariate CvM-distance is a member of the generalized energy distance by the use of the angular distance and thus also a member of the MMD. Let $\mathcal{M}_X$ and $\mathcal{M}_Y$ be the support of X and Y respectively and let $\mathcal{M} = \mathcal{M}_X \cup \mathcal{M}_Y \subseteq \mathbb{R}^d$. Then we define the *angular distance* as follows:

Definition 4.2 (Angular distance)**.** *Let Z^* be a random vector having mixture distribution $(1/2)P_X + (1/2)P_Y$. For $z, z' \in \mathcal{M}$, denote the scaled angle between $z - Z^*$ and $z' - Z^*$ by*

$$\rho_{\mathrm{Angle}}(z, z'; Z^*) = \frac{1}{\pi} \mathsf{Ang}\left(z - Z^*, z' - Z^*\right).$$

The angular distance is defined as the expected value of the scaled angle:

$$\rho_{\mathrm{Angle}}(z, z') = \mathbb{E}\left[\rho_{\mathrm{Angle}}(z, z'; Z^*)\right]. \tag{4.18}$$

As shown in Appendix C.5.6, ρ_{Angle} is a metric of negative type defined on $\mathcal{M}$. With this key property and the identity given in the next proposition, we may conclude that the multivariate CvM-distance is a special case of the generalized energy distance based on the angular distance.

Proposition 4.2 (Another view of the CvM-distance)**.** *Let us consider the angular distance defined in (4.18). Then*

$$2W_d^2 = 2\mathbb{E}\left[\rho_{\mathrm{Angle}}(X_1, Y_1)\right] - \mathbb{E}\left[\rho_{\mathrm{Angle}}(X_1, X_2)\right] - \mathbb{E}\left[\rho_{\mathrm{Angle}}(Y_1, Y_2)\right].$$

Remark 4.6. The angular distance can be generalized by taking the expectation with respect to a different measure. For instance, when the expectation is taken with respect to Lebesgue measure, the generalized

angular distance is proportional to the Euclidean distance (see Appendix C.5.7). The main difference between the Euclidean distance and the proposed angular distance is that the latter takes into account information from the underlying distribution and is less sensitive to outliers. In this respect, the introduced angular distance can be viewed as a robust alternative for the Euclidean distance.

Remark 4.7. As one of the reviewers pointed out, there might be several ways to enhance the power of the proposed test by modifying the multivariate CvM-distance. For instance, by the characteristic property of W_d^2, it can be seen that H_0 holds if and only if $\mathsf{T}_{xy} = \mathsf{T}_{xx}$ and $\mathsf{T}_{xy} = \mathsf{T}_{yy}$ where $\mathsf{T}_{xy} = \mathbb{E}[\rho_{\text{Angle}}(X_1, Y_1)]$, $\mathsf{T}_{xx} = \mathbb{E}[\rho_{\text{Angle}}(X_1, X_2)]$ and $\mathsf{T}_{yy} = \mathbb{E}[\rho_{\text{Angle}}(Y_1, Y_2)]$. Motivated by this observation, another test statistic can be introduced based on an estimate of $(\mathsf{T}_{xy} - \mathsf{T}_{xx})^2 + (\mathsf{T}_{xy} - \mathsf{T}_{yy})^2$ (and other variants are possible, see Appendix C.5.4). As demonstrated in Appendix C.6, the test based on this new statistic tends to be more sensitive to scale differences than the CvM test.

4.7 Other Multivariate Extensions via Projection-Averaging

The projection-averaging approach used for the multivariate CvM-statistic is general and can be applied to many other univariate robust statistics. In this section and also Appendix C.5.8, we illustrate the utility of the projection-averaging approach by considering several examples including Kendall's tau, Spearman's rho and the sign covariance (Bergsma and Dassios, 2014). Let us begin with one-sample and two-sample robust statistics. Given a pair of random variables (X, Y), denote the difference between two random variables by $Z = X - Y$. The univariate sign test statistic is an estimate of $\mathsf{T}_{\text{sign}} := \mathbb{P}(Z > 0) - 1/2$ and it is used to test whether

$$H_0 : \mathbb{P}(Z > 0) = 1/2 \quad \text{versus} \quad H_1 : \mathbb{P}(Z > 0) \neq 1/2.$$

The projection-averaging technique extends T_{sign} to a multivariate case as follows:

Proposition 4.3 (One-sample sign test statistic)**.** *For i.i.d. random vectors Z_1, Z_2 from a multivariate distribution P_Z where $Z \in \mathbb{R}^d$, the projection-averaging approach generalizes T_{sign} as*

$$\int_{\mathbb{S}^{d-1}} \left(\mathbb{P}(\beta^\top Z_1 > 0) - \frac{1}{2} \right)^2 d\lambda(\beta) = \frac{1}{4} - \frac{1}{2\pi} \mathbb{E}\left[\mathsf{Ang}\left(Z_1, Z_2 \right) \right]. \tag{4.19}$$

Given univariate two samples $\mathcal{X}_m = \{X_1, \ldots, X_m\}$ and $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$, the Wilcoxon–Mann–Whitney test is designed for testing whether

$$H_0 : \mathbb{P}(X > Y) = 1/2 \quad \text{versus} \quad H_1 : \mathbb{P}(X > Y) \neq 1/2,$$

based on an estimate of $\mathsf{T}_{\mathrm{WMW}} := \mathbb{P}(X > Y) - 1/2$. The next proposition extends $\mathsf{T}_{\mathrm{WMW}}$ to a multivariate case via projection-averaging.

Proposition 4.4 (Two-sample Wilcoxon–Mann–Whitney test statistic)**.** *Let* $X_1, X_2 \overset{i.i.d.}{\sim} P_X$ *and, independently,* $Y_1, Y_2 \overset{i.i.d.}{\sim} P_Y$ *where* $X_1, Y_1 \in \mathbb{R}^d$*. The projection-averaging approach generalizes* $\mathsf{T}_{\mathrm{WMW}}$ *as*

$$\int_{\mathbb{S}^{d-1}} \left(\mathbb{P}(\beta^\top X_1 > \beta^\top Y_1) - \frac{1}{2} \right)^2 d\lambda(\beta) = \frac{1}{4} - \frac{1}{2\pi} \mathbb{E}\left[\mathsf{Ang}\left(X_1 - Y_1, X_2 - Y_2 \right) \right]. \tag{4.20}$$

Remark 4.8. The first order Taylor approximation of the inverse cosine function shows that the representations given on the right-side of (4.19) and (4.20) are related to the spatial sign-statistics introduced by Wang et al. (2015) and Chakraborty and Chaudhuri (2017), respectively. In fact, when U-statistics are used to estimate (4.19) and (4.20), the projection-averaging statistics and the spatial sign-statistics are asymptotically equivalent under some regularity conditions (see Appendix C.5.3 for details).

The same technique can be further applied to some robust statistics for independence testing. To test for independence between two random variables, Kendall's tau statistic is given as an estimate of $\tau := 4\mathbb{P}\left(X_1 < X_2, Y_1 < Y_2\right) - 1$. We present a multivariate extension of τ as follows:

Theorem 4.12 (Kendall's tau)**.** *For i.i.d. pairs of random vectors* (X_1, Y_1), $\ldots, (X_4, Y_4)$ *from a joint distribution* P_{XY} *where* $X \in \mathbb{R}^p$ *and* $Y \in \mathbb{R}^q$*, the multivariate extension of* τ *via projection-averaging is given by*

$$\begin{aligned} &\int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{q-1}} \left[4\mathbb{P}\left(\alpha^\top (X_1 - X_2) < 0, \beta^\top (Y_1 - Y_2) < 0 \right) - 1 \right]^2 d\lambda(\alpha) d\lambda(\beta) \\ = \;& \mathbb{E}\left[\left(2 - \frac{2}{\pi} \mathsf{Ang}\left(X_1 - X_2, X_3 - X_4 \right) \right) \cdot \left(2 - \frac{2}{\pi} \mathsf{Ang}\left(Y_1 - Y_2, Y_3 - Y_4 \right) \right) \right] - 1. \end{aligned}$$

Recently, Bergsma and Dassios (2014) introduced a modification of Kendall's tau, which is zero if and only if random variables are independent under some mild conditions. Let us denote the univariate Bergsma–Dassios sign covariance by

$$\tau^* = \mathbb{E}\left[a_{\mathrm{sign}}(X_1, X_2, X_3, X_4) \cdot a_{\mathrm{sign}}(Y_1, Y_2, Y_3, Y_4) \right], \tag{4.21}$$

with $a_{\mathrm{sign}}(z_1, z_2, z_3, z_4) = \mathrm{sign}\left(|z_1 - z_2| + |z_3 - z_4| - |z_1 - z_3| - |z_2 - z_4| \right)$. Motivated by the projection-averaging approach, we propose the multivariate τ^* as follows:

Definition 4.3 (Multivariate τ^*)**.** *Suppose* $(X_1, Y_1), \ldots, (X_4, Y_4)$ *are i.i.d. random vectors from a joint distribution* P_{XY} *where* $X \in \mathbb{R}^p$ *and* $Y \in \mathbb{R}^q$*. We define the multivariate* τ^* *by*

$$\tau^*_{p,q} = \int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{q-1}} \mathbb{E}\big[a_{\text{sign}}(\alpha^\top X_1, \alpha^\top X_2, \alpha^\top X_3, \alpha^\top X_4) \times a_{\text{sign}}(\beta^\top Y_1, \beta^\top Y_2, \beta^\top Y_3, \beta^\top Y_4)\big] d\lambda(\alpha) d\lambda(\beta).$$

Since the kernel of τ^* is sign-invariant, i.e. $a_{\text{sign}}(z_1, z_2, z_3, z_4) = a_{\text{sign}}(-z_1, -z_2, -z_3, -z_4)$, it is easy to see that $\tau^*_{p,q}$ becomes the univariate τ^* when $p = q = 1$. Also note that since X and Y are independent if and only if $\alpha^\top X$ and $\beta^\top Y$ are independent for all $\alpha \in \mathbb{S}^{p-1}$ and $\beta \in \mathbb{S}^{q-1}$, the characteristic property of $\tau^*_{p,q}$ follows by that of the univariate τ^*.

Next we present a closed-form expression for $\tau^*_{p,q}$. For non-zero $U_1, U_2, U_3 \in \mathbb{R}^d$, let us define $g_d(U_1, U_2, U_3)$ and $h_d(Z_1, Z_2, Z_3, Z_4)$ by

$$g_d(U_1, U_2, U_3) = \frac{1}{2} - \frac{1}{4\pi} \left[\mathsf{Ang}\left(U_1, U_2\right) + \mathsf{Ang}\left(U_1, U_3\right) + \mathsf{Ang}\left(U_2, U_3\right)\right]$$

and

$$\begin{aligned} h_d(Z_1, Z_2, Z_3, Z_4) \;=\; & g_d(Z_1 - Z_2, Z_2 - Z_3, Z_3 - Z_4) + g_d(Z_2 - Z_1, Z_1 - Z_3, Z_3 - Z_4) \\ & + g_d(Z_1 - Z_2, Z_2 - Z_4, Z_4 - Z_3) + g_d(Z_2 - Z_1, Z_1 - Z_4, Z_4 - Z_3). \end{aligned}$$

We note that, in contrast to other applications, Lemma 4.0.2 is not enough to have an expression for $\tau^*_{p,q}$ without involving integrations over the unit sphere. To this end, we generalize Lemma 4.0.2 with three indicator functions (see Lemma C.1.8) and present an alternative expression for $\tau^*_{p,q}$ as follows.

Theorem 4.13 (Closed-form expression for $\tau^*_{p,q}$)**.** *For i.i.d. random vectors* $(X_1, Y_1), \ldots, (X_4, Y_4)$ *from a joint distribution* P_{XY} *where* $X \in \mathbb{R}^p$ *and* $Y \in \mathbb{R}^q$*,* $\tau^*_{p,q}$ *can be written as*

$$\begin{aligned} \tau^*_{p,q} \;=\; & \mathbb{E}\left[h_p(X_1, X_2, X_3, X_4) \cdot h_q(Y_1, Y_2, Y_3, Y_4)\right] + \mathbb{E}\left[h_p(X_1, X_2, X_3, X_4) \cdot h_q(Y_3, Y_4, Y_1, Y_2)\right] \\ & - 2\mathbb{E}\left[h_p(X_1, X_2, X_3, X_4) \cdot h_q(Y_1, Y_3, Y_2, Y_4)\right]. \end{aligned}$$

Theorem 4.13 leads to a straightforward empirical estimate of $\tau^*_{p,q}$ based on a U-statistic. This is also true for the other multivariate generalizations introduced in this section and the supplementary material (Appendix C.5.8). Using these estimates, some theoretical and empirical properties of the proposed measures can be further investigated. These topics are reserved for future work.

4.8 Simulations

In this section, we report numerical results to support the argument in Section 4.5 as well as to compare the performance of the CvM test with other competing nonparametric tests against heavy-tailed alternatives. Along with the energy, MMD, NN, FR and BG tests described before, we consider the cross-match test (Rosenbaum, 2005), the multivariate run test (Biswas et al., 2014), the modified k-NN test (Mondal et al., 2015) and the ball divergence test (Pan et al., 2018) for comparison. We refer to them as the CM test, run test, MBG test and ball test, respectively. In our simulations, we used the Gaussian kernel with the median heuristic (Gretton et al., 2012) for the MMD test and we set the number of nearest neighbors as $k = 3$ for both NN test and MBG test. Since finding the shortest Hamiltonian path for the run test is NP-complete, we employed Kruskal's algorithm (Kruskal, 1956) as suggested by Biswas et al. (2014).

Throughout our experiments, the significance level was set at 0.05 and the permutation procedure was used to determine the p-value of each test with 200 permutations as in Remark 4.2. The simulations were repeated 500 times to approximate the power of different tests. We set the sample size and the dimension by $m, n = 20$ and $d = 200$ for the balanced cases and by $m = 35, n = 5$ and $d = 200$ for the imbalanced cases.

First, we consider several examples where the powers of the five tests (CvM, energy, MMD, CQ and WMW tests) in Section 4.5 are approximately equivalent to each other. Specifically we use multivariate normal distributions with different means

$$\mu^{(0)} = (0, \ldots, 0)^\top, \quad \mu^{(1)} = (0.15, \ldots, 0.15)^\top \quad \text{and}$$
$$\mu^{(2)} = \sqrt{0.045}(\underbrace{1, \ldots, 1}_{d/2 \text{ elements}} , \underbrace{0, \ldots, 0}_{d/2 \text{ elements}})^\top$$

and covariance matrices:

1. Identity matrix (denoted by I) where $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0$ for $i \neq j$.
2. Banded matrix (denoted by Σ_{Band}) where $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.6$ for $|i - j| = 1$, $\sigma_{i,j} = 0.3$ for $|i - j| = 2$ and $\sigma_{i,j} = 0$ otherwise.
3. Autocorrelation matrix (denoted by Σ_{Auto}) where $\sigma_{i,i} = 1$ and $\sigma_{i,j} = 0.2^{|i-j|}$ when $i \neq j$.
4. Block diagonal matrix (denoted by Σ_{Block}) where the 5×5 main diagonal blocks $\mathbf{A}$ are defined by $a_{i,i} = 1$ and $a_{i,j} = 0.2$ when $i \neq j$, and the off-diagonal blocks are zeros.

Then we generate random samples from $X \sim N(\mu^{(0)}, \Sigma)$ and either $Y \sim N(\mu^{(1)}, \Sigma)$ or $Y \sim N(\mu^{(2)}, \Sigma)$. The results are summarized in Table 4.1. As can be seen from the table, the empirical powers of the considered tests are very close under the given setting, which supports our theoretical results in Section 4.5. We also observe that the other nonparametric tests, not considered in Section 4.5, are significantly less powerful than the proposed test in all normal location alternatives.

Table 4.1: Empirical power of the considered tests against the normal location models at $\alpha = 0.05$.

	I_d		Σ_{Band}		Σ_{Block}		Σ_{Auto}	
$m = 20, n = 20$	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(1)}$	$\mu^{(2)}$	$\mu^{(1)}$	$\mu^{(2)}$
CvM	**0.662**	**0.646**	**0.418**	**0.406**	**0.572**	**0.584**	**0.452**	**0.442**
Energy	**0.656**	**0.650**	**0.420**	**0.408**	**0.576**	**0.584**	**0.452**	**0.444**
MMD	**0.658**	**0.638**	**0.412**	**0.398**	**0.568**	**0.570**	**0.458**	**0.444**
CQ	**0.656**	**0.650**	**0.416**	**0.412**	**0.578**	**0.580**	**0.454**	**0.448**
WMW	**0.668**	**0.646**	**0.420**	**0.402**	**0.568**	**0.580**	**0.458**	**0.444**
NN	0.288	0.288	0.164	0.154	0.242	0.238	0.176	0.174
FR	0.168	0.170	0.090	0.084	0.158	0.116	0.112	0.088
MBG	0.050	0.050	0.050	0.052	0.048	0.044	0.060	0.046
Ball	0.240	0.254	0.186	0.198	0.262	0.250	0.216	0.226
CM	0.042	0.054	0.028	0.040	0.052	0.050	0.038	0.034
BG	0.070	0.060	0.074	0.074	0.074	0.078	0.084	0.078
Run	0.160	0.153	0.101	0.105	0.146	0.128	0.110	0.102

In our second experiment, we consider several examples where the moment conditions are not satisfied. We focus on random samples generated from multivariate Cauchy distributions. Let Cauchy(γ, s) refer to the univariate Cauchy distribution where γ, s are the location parameter and the scale parameter, respectively. Let $X = (X^{(1)}, \ldots, X^{(d)})$ and $Y = (Y^{(1)}, \ldots, Y^{(d)})$ be random vectors where $X^{(i)} \overset{i.i.d.}{\sim}$ Cauchy$(0, 1)$ and $Y^{(i)} \overset{i.i.d.}{\sim}$ Cauchy(γ, s) for $i = 1, \ldots, d$. We first consider location differences where γ is not zero but the scale parameters are identical, i.e. $s = 1$. Similarly, we consider scale differences where the scale parameter s changes, but the location parameters are identical, i.e. $\gamma = 0$.

From the results presented in Table 4.2 and Table 4.3, it is seen that, unlike the multivariate normal cases, there are significant differences between power performance among CvM, energy, MMD, CQ and WMW tests. In particular, the tests based on the energy, MMD and CQ statistics have relatively low power against the heavy-tail location alternatives, whereas the tests based on the CvM and WMW statistics show better performance than the others. Turning to the scale problems, it can be seen that the CQ and WMW tests are not sensitive to detect scale differences, which makes sense because they are specifically designed for location problems. On the other hand, the CvM, energy and MMD tests perform reasonably well in these alternatives. Among the omnibus nonparametric tests, the MMD, energy and ball tests have competitive power against the scale differences, but not against the location differences in general. The MBG test is only powerful against the scale differences where the sample sizes are balanced. The CM and run tests are uniformly outperformed by the CvM test under all scenarios. The NN and FR tests perform strongly against

Table 4.2: Empirical power of the considered tests against multivariate Cauchy distributions with $m = n = 20$ at $\alpha = 0.05$ where γ, s represent the location and scale parameter, respectively. The three highest power estimates in each column are highlighted in boldface.

	Location				*Scale*			
$m = 20, n = 20$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$	$\gamma = 5$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
CvM	0.124	0.252	0.596	0.842	**0.560**	**0.926**	**0.988**	**1.000**
Energy	0.060	0.066	0.102	0.134	0.316	0.602	0.766	0.866
MMD	0.056	0.064	0.110	0.162	0.448	0.772	0.890	0.970
CQ	0.138	0.268	0.360	0.456	0.046	0.070	0.042	0.068
WMW	**0.324**	**0.698**	**0.912**	**0.988**	0.052	0.064	0.062	0.056
NN	**0.288**	**0.662**	**0.884**	**0.976**	0.214	0.194	0.256	0.224
FR	**0.178**	**0.462**	**0.706**	**0.888**	0.028	0.034	0.048	0.036
MBG	0.060	0.044	0.050	0.074	**0.564**	**0.904**	**0.964**	**0.992**
Ball	0.064	0.064	0.076	0.098	**0.606**	**0.936**	**0.994**	**1.000**
CM	0.030	0.078	0.128	0.226	0.056	0.170	0.334	0.490
BG	0.048	0.038	0.048	0.040	0.238	0.394	0.560	0.632
Run	0.059	0.129	0.274	0.422	0.220	0.506	0.767	0.864

Table 4.3: Empirical power of the considered tests against multivariate Cauchy distributions with $m = 35$ and $n = 5$ at $\alpha = 0.05$ where γ, s represent the location and scale parameter, respectively. The three highest power estimates in each column are highlighted in boldface.

	Location				*Scale*			
$m = 35, n = 5$	$\gamma = 5$	$\gamma = 6$	$\gamma = 7$	$\gamma = 8$	$s = 3$	$s = 4$	$s = 5$	$s = 6$
CvM	**0.340**	**0.498**	**0.652**	**0.758**	**0.570**	**0.806**	**0.928**	**0.952**
Energy	0.110	0.146	0.212	0.262	**0.436**	**0.632**	**0.794**	**0.858**
MMD	0.108	0.148	0.192	0.240	**0.552**	**0.808**	**0.926**	**0.968**
CQ	**0.284**	**0.380**	0.454	0.544	0.178	0.210	0.262	0.290
WMW	**0.796**	**0.890**	**0.942**	**0.960**	0.110	0.126	0.134	0.148
NN	0.144	0.294	0.376	0.558	0.118	0.150	0.154	0.182
FR	0.226	0.360	**0.464**	**0.588**	0.078	0.092	0.104	0.112
MBG	0.010	0.000	0.008	0.000	0.092	0.130	0.176	0.214
Ball	0.072	0.088	0.098	0.122	0.238	0.406	0.594	0.762
CM	0.082	0.176	0.190	0.262	0.030	0.080	0.092	0.126
BG	0.058	0.052	0.058	0.052	0.320	0.386	0.506	0.514
Run	0.088	0.150	0.198	0.228	0.106	0.174	0.248	0.326

the location alternatives especially for the balanced case, but not against the scale alternatives. When the sample sizes are unbalanced, the performance of the NN and FR tests are degraded a little bit, which can be explained by Chen et al. (2013) and Chen et al. (2018). The CvM test, on the other hand, performs consistently well against the heavy-tail location and scale alternatives and its performance appears immune to the sample proportion.

In summary, the proposed test has almost identical power as the high-dimensional mean tests against the light-tail location alternatives, whereas it outperforms many popular nonparametric competitors under the heavy-tail location and scale alternatives. More simulation results in both high and low dimensions can be found in Appendix C.6 of the supplemental article.

4.9 Concluding Remarks

In this work, we extended the univariate Cramér-von Mises statistic for two-sample testing to the multivariate case using projection-averaging. The proposed statistic has a straightforward calculation formula in arbitrary dimensions and the resulting test has good statistical properties. Throughout this chapter, we demonstrated its robustness, minimax rate optimality and high-dimensional power properties. In addition, we applied the same projection technique to other robust statistics and presented their multivariate extensions.

Beyond nonparametric testing problems, we believe that our approach can be used for other problems. For example, our work can be viewed as an application of the angular distance to the two-sample problem. The angular distance is closely connected to the Euclidean distance (Remark 4.6) but is more robust to outliers by incorporating information from the underlying distribution. Given that the use of distances is of fundamental importance in many statistical applications (including clustering, classification and regression), we expect that the angular distance can be applied to other statistical problems as a robust alternative for the Euclidean distance.

Chapter 5

Comparing a Large Number of Multivariate Distributions

This chapter is adapted from my work supervised by Sivaraman Balakrishnan and Larry Wasserman. This work is available on ArXiv (Kim, 2019).

5.1 Introduction

Let $P_1, \ldots, P_K$ be probability distributions defined on a common measurable space $(\mathcal{X}, \mathcal{B})$ for $K \geq 2$. The K-sample problem is concerned with testing the null hypothesis $H_0 : P_1 = \cdots = P_K$ against the alternative hypothesis $H_1 : P_i \neq P_j$ for some $i, j \in \{1, \ldots, K\}$. This fundamental problem of comparing multiple distributions is a classical topic in statistics with a wide range of applications (Thas, 2010; Chen and Pokojovy, 2018, for reviews). Despite its long history, previous approaches to the K-sample problem have several limitations. First, many methods are limited to dealing with univariate data. For instance, Kiefer (1959) proposes the K-sample analogues of the Kolmogorov–Smirnov and Cramér-Von Mises tests. Scholz and Stephens (1987) generalize the Anderson-Darling test (Anderson and Darling, 1952) to the K-sample case. These approaches are based on empirical distribution functions and are not easily extendable to multivariate data. Some other references that are restricted to the univariate K-sample problem include Conover (1965); Zhang and Wu (2007); Wyłupek (2010); Quessy and Éthier (2012); Lemeshko and Veretelnikova (2018). Second, most research in this area has been carried out under classical asymptotic regimes where the sample size goes to infinity but the number of distributions is fixed (e.g., Burke, 1979; Bouzebda et al., 2011; Hušková and Meintanis, 2008; Martínez-Camblor et al., 2008; Jiang et al., 2015; Mukhopadhyay and Wang, 2018; Sosthene et al., 2018). Clearly this classical asymptotic analysis is not appropriate for a dataset with large K and it only provides a narrow picture of the behavior of a test. To

the best of our knowledge, Zhan and Hart (2014) is the only study in the literature that considers large K. However, their analysis is limited to univariate data with fixed sample size. Third, recent developments on the multivariate K-sample problem are largely built upon an average difference between distributions (Bouzebda et al., 2011; Hušková and Meintanis, 2008; Rizzo and Székely, 2010; Zhan and Hart, 2014; Mukhopadhyay and Wang, 2018; Sosthene et al., 2018). It is well-known that the test based on an average-type test statistic tends to be powerful against dense alternatives in which many of $P_1, \ldots, P_K$ are different to each other. On the other hand, it tends to suffer from low power against sparse alternatives where only a few of $P_1, \ldots, P_K$ are different from the others. Recently, sparse alternatives have been motivated by numerous applications such as DNA microarray analysis and anomaly detection where there are a small number of treatments that can actually contribute response variables. These applications have led to recent developments of tests tailored to sparse alternatives in the context of testing a high-dimensional vector (Jeng et al., 2010; Fan et al., 2015; Liu and Li, 2020), two-sample mean or covariance testing (Cai et al., 2013, 2014; Cai and Xia, 2014), analysis of variance (Arias-Castro et al., 2011; Cai and Xia, 2014) and independence testing (Han et al., 2017). To our knowledge, however, a multivariate K-sample test specifically designed for sparse alternatives is not available in the current literature.

In this study, we propose a new K-sample test that addresses the aforementioned limitations of the previous approaches. More specifically, we introduce a K-sample test based on the kernel mean embedding method that has been successfully applied to multivariate hypothesis testing. Our test statistic is defined as the maximum of pairwise maximum mean discrepancies (Gretton et al., 2007, 2012), which leads to a powerful test against sparse alternatives. Throughout this chapter, we investigate statistical properties of the proposed test under the asymptotic regime where both the sample size and the number of distributions tend to infinity. Below, we summarize our main findings and contributions.

- **Limiting null distribution:** By building on Drton et al. (2018), we develop Cramér-type moderate deviation for degenerate two-sample V-statistics. Based on this result, we study the limiting distribution of the proposed test statistic when the sample size and the number of distributions increase simultaneously. In particular, we show the test statistic converges to a Gumbel distribution under some appropriate conditions.

- **Concentration inequality under permutations:** We demonstrate the usefulness of Bobkov's inequality (Bobkov, 2004) in studying a concentration inequality for the permuted test statistic. By applying his result, we derive an exponential concentration inequality for the proposed test statistic under permutations. In contrast to usual Hoeffding or Bernstein-type inequalities, the developed inequality relies solely on completely known and easily computable quantities without any moment assumption.

- **Uniform consistency of the permutation test:** Leveraging the developed concentration inequality for the permuted statistic, we prove the uniform consistency of the permutation test over the class of sparse alternatives. Under some regularity conditions, we also show that the power of the permutation test cannot be improved from a minimax point of view.
- **Empirical power comparison against sparse alternatives:** A simulation study is conducted to compare the performance of the proposed maximum-type test with the existing average-type tests in the literature. The simulation results show that the proposed test consistently outperforms the average-type tests against different sparse alternatives.

Outline. This chapter is organized as follows. In Section 5.2, we briefly review the maximum mean discrepancy and introduce our test statistic. Section 5.3 studies the limiting distribution of the proposed test statistic when the sample size and the number of distributions tend to infinity simultaneously. Section 5.4 formally introduces permutation procedures. In Section 5.5, we provide an exponential concentration inequality for the proposed test statistic under permutations. Section 5.6 investigates the power of the proposed test and proves its optimality property against sparse alternatives. In Section 8.9, we demonstrate the finite-sample performance of the proposed approach via simulations. Finally, Section 5.8 concludes this chapter and discusses future work. The proofs not presented in the main text can be found in Appendix D.1.

5.2 Test Statistic

We start with a brief overview of the maximum mean discrepancy proposed by Gretton et al. (2007, 2012). Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) on $\mathcal{X}$ with a reproducing kernel $h : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. For two functions $f, g \in \mathcal{H}$, we write the inner product on $\mathcal{H}$ by $\langle f, g \rangle_{\mathcal{H}}$ and the associated norm by $\|f\|_{\mathcal{H}}$. Given a probability distribution P, the kernel mean embedding of P is given by $\mu_h(P) = \mathbb{E}_{X \sim P}[h(X, \cdot)]$. Using the feature map $\psi : \mathcal{X} \mapsto \mathcal{H}$, which satisfies $h(x, y) = \langle \psi(x), \psi(y) \rangle_{\mathcal{H}}$, the kernel mean embedding can also be written as $\mathbb{E}_{X \sim P}[\psi(X)]$ (see e.g. Muandet et al., 2016, for details). We now provide the definition of the maximum mean discrepancy (MMD) associated with kernel h.

Definition 5.1 (Maximum mean discrepancy)**.** *Given two probability distributions, say P_1 and P_2, such that $\mathbb{E}_{X_1 \sim P_1} \|\psi(X_1)\|_{\mathcal{H}} < \infty$ and $\mathbb{E}_{X_2 \sim P_2} \|\psi(X_2)\|_{\mathcal{H}} < \infty$, the maximum mean discrepancy is defined as the RKHS norm of the difference between $\mu_h(P_1)$ and $\mu_h(P_2)$, i.e.*

$$\mathcal{V}_h(P_1, P_2) = \|\mu_h(P_1) - \mu_h(P_2)\|_{\mathcal{H}}.$$

It has been shown that when kernel h is characteristic (see e.g., Fukumizu et al., 2008; Sriperumbudur et al., 2011), the MMD becomes zero *if and only if* $P_1 = P_2$. Some examples of characteristic kernels include Gaussian and Laplace kernels on $\mathcal{X} = \mathbb{R}^d$. This characteristic property allows to have a consistent two-sample test against any fixed alternatives. For general K-sample cases, we consider the maximum of pairwise maximum mean discrepancies as our metric, i.e.

$$\mathcal{V}_{h,\max}(P_1, \ldots, P_K) = \max_{1 \le k < l \le K} \|\mu_h(P_k) - \mu_h(P_l)\|_{\mathcal{H}}.$$

Hence as long as h is characteristic, it is clear to see that $\mathcal{V}_{h,\max}(P_1, \ldots, P_K)$ is zero *if and only if* $P_1 = \cdots = P_K$.

Suppose that we observe identically distributed samples $X_{1,k}, \ldots, X_{n_k,k} \sim P_k$ for each $k = 1, \ldots, K$ and assume that the samples are mutually independent. We propose our test statistic defined as a plug-in estimator of $\mathcal{V}_{h,\max}$:

$$\widehat{\mathcal{V}}_{h,\max} = \max_{1 \le k < l \le K} \left\| \frac{1}{n_k} \sum_{i_1=1}^{n_k} \psi(X_{i_1,k}) - \frac{1}{n_l} \sum_{i_2=1}^{n_l} \psi(X_{i_2,l}) \right\|_{\mathcal{H}}.$$

In practice, the test statistic can be computed in a straightforward manner based on the kernel trick (e.g. Lemma 6 of Gretton et al., 2012):

$$\widehat{\mathcal{V}}_{h,\max} = \max_{1 \le k < l \le K} \left\{ \frac{1}{n_k^2} \sum_{i_1,i_2=1}^{n_k} h(X_{i_1,k}, X_{i_2,k}) + \frac{1}{n_l^2} \sum_{i_1,i_2=1}^{n_l} h(X_{i_1,l}, X_{i_2,l}) - \frac{2}{n_k n_l} \sum_{i_1=1}^{n_k} \sum_{i_2=1}^{n_l} h(X_{i_1,k}, X_{i_2,l}) \right\}^{1/2}.$$

Throughout this chapter, we denote the pooled samples by $\{Z_1, \ldots, Z_N\} = \{X_{1,1}, \ldots, X_{n_K,K}\}$ where $N = \sum_{k=1}^{K} n_k$.

5.3 Limiting distribution

Given the test statistic, our next step is to determine a critical value of the test with correct size α and good power properties. A common way of calibrating the critical value is based on the limiting null distribution of the test statistic. In this asymptotic approach, the critical value is set to be the $1 - \alpha$ quantile of the limiting null distribution and the null hypothesis is rejected when the test statistic exceeds the critical value. The purpose of this section is to demonstrate the difficulty of implementing this asymptotic-based test in our setting. In particular, we show that $\widehat{\mathcal{V}}_{h,\max}$ converges to a Gumbel distribution with a potentially infinite number of unknown parameters under certain conditions. Unfortunately, it is by no means trivial to consistently estimate these infinite nuisance parameters. Furthermore, it is well-known that a maximum-type statistic converges slowly to its limiting distribution (e.g. Hall, 1991), which also makes the asymptotic test

less attractive in practice. These limitations motivate us to delve into the permutation approach later in Section 5.4–5.6.

5.3.1 Cramér-type moderate deviation

In order to derive the limiting distribution of the maximum of pairwise MMD statistics, it is important to understand the tail behavior of the two-sample MMD statistic. The main tool to this end is Cramér-type moderate deviation for degenerate two-sample V-statistics that we will develop in this subsection. Our result largely builds upon Cramér-type moderate deviation for degenerate one-sample U-statistics recently presented by Drton et al. (2018).

Let us start with some notation and assumptions. For notational convenience, we write the MMD statistic between P_1 and P_2 as

$$\widehat{\mathcal{V}}_{12}^2 = \frac{1}{n_1^2} \sum_{i_1,i_2=1}^{n_1} h(X_{i_1,1}, X_{i_2,1}) + \frac{1}{n_2^2} \sum_{i_1,i_2=1}^{n_2} h(X_{i_1,2}, X_{i_2,2}) - \frac{2}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} h(X_{i_1,1}, X_{i_2,2}).$$

By defining $h^*(x_1, x_2; y_1, y_2) := h(x_1, x_2) + h(y_1, y_2) - h(x_1, y_1)/2 - h(x_1, y_2)/2 - h(x_2, y_1)/2 - h(x_2, y_2)/2$, the MMD statistic can also be written in the form of a two-sample V-statistic

$$\widehat{\mathcal{V}}_{12}^2 = \frac{1}{n_1^2 n_2^2} \sum_{i_1,i_2=1}^{n_1} \sum_{j_1,j_2=1}^{n_2} h^*(X_{i_1,1}, X_{i_2,1}; X_{j_1,2}, X_{j_2,2}). \tag{5.1}$$

Under the null hypothesis, the considered V-statistic is *degenerate* meaning that the conditional expectation of $h^*(X_{i_1,1}, X_{i_2,1}; X_{j_1,2}, X_{j_2,2})$ given any one of $X_{i_1,1}, X_{i_2,1}, X_{j_1,2}, X_{j_2,2}$ has zero variance whenever $i_1 \neq i_2$ and $j_1 \neq j_2$.

Let X_1, X_2 be independent random vectors from P_1. We then define the centered kernel

$$\overline{h}(x_1, x_2) := h(x_1, x_2) - \mathbb{E}[h(x_1, X_2)] - \mathbb{E}[h(X_1, x_2)] + \mathbb{E}[h(X_1, X_2)],$$

which satisfies $\mathbb{E}[\overline{h}(X_1, X_2)] = 0$ and $\mathbb{E}[\overline{h}(x_1, X_2)] = 0$ almost surely. Under the finite second moment condition of the centered kernel, i.e. $\mathbb{E}[\{\overline{h}(X_1, X_2)\}^2] < \infty$, we may write

$$\overline{h}(x_1, x_2) = \sum_{v=1}^{\infty} \lambda_v \varphi_v(x_1) \varphi_v(x_2), \tag{5.2}$$

where $\{\lambda_v\}_{v=1}^{\infty}$ and $\{\varphi_v(\cdot)\}_{v=1}^{\infty}$ are the eigenvalues and eigenfunctions of the integral equation $\mathbb{E}[\overline{h}(x_1, X_2)\varphi_v(X_2)] = \lambda_v \varphi_v(x_1)$ (e.g. page 80 of Lee, 1990).

To facilitate the analysis, we make the following assumptions regarding the kernel function.

(A1). Assume that $\mathbb{E}[|\overline{h}(X_1, X_1)|] < \infty$.

(A2). Suppose that $\overline{h}(x_1, x_2)$ admits the decomposition in (5.2) with $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. For all $u, v \in \mathbb{S}^{T-1} := \{x \in \mathbb{R}^T : \|x\|_2 = 1\}$ where $\|\cdot\|_2$ is Euclidean norm in $\mathbb{R}^T$ and any positive integer T, assume that there exists a constant $\eta > 0$ independent of T such that

$$\mathbb{E}\big[\big|\{\varphi_{1\cdots T}(X_1)^\top u\}^2\{\varphi_{1\cdots T}(X_1)^\top v\}^{m-2}\big|\big] \leq \eta^m m^{m/2}, \tag{5.3}$$

where $\varphi_{1\cdots T}(X_1) := (\varphi_1(X_1), \ldots, \varphi_T(X_1))^\top$ and $m = 3, 4, \ldots$

It is worth noting that the given conditions are more general than those used in Drton et al. (2018). Specifically, Drton et al. (2018) assume that the kernel h and its eigenfunctions are uniformly bounded. Clearly, **(A1)** and **(A2)** are fulfilled under their boundedness assumptions. We also note that $\overline{h}(x_1, x_2)$ is a valid positive definite kernel (Sejdinovic et al., 2013), which yields $\{\overline{h}(x_1, x_2)\}^2 \leq \overline{h}(x_1, x_1)\overline{h}(x_2, x_2)$. Hence, the second moment condition $\mathbb{E}[\{\overline{h}(X_1, X_2)\}^2] < \infty$ is also satisfied under **(A1)**. Finally, the multivariate moment condition (5.3) implies that individual eigenfunctions are sub-Gaussian (e.g. Vershynin, 2018).

Under the given conditions, we present Cramér-type moderate deviation for the two-sample degenerate V-statistic described in (5.1). The proof of the following theorem can be found in Appendix D.1.

Theorem 5.1 (Cramér-type moderate deviation)**.** *Suppose that **(A1)** and **(A2)** are fulfilled. Assume that there exists a constant $C_1 \geq 1$ such that $C_1^{-1} \leq n_1/n_2 \leq C_1$ and n_1/N converges to a constant as $N := n_1 + n_2 \to \infty$. Then under the null hypothesis $P_1 = P_2$, we have*

$$\frac{\mathbb{P}(n_1 n_2 \widehat{\mathcal{V}}_{12}^2/N \geq x)}{\mathbb{P}\left(\sum_{v=1}^{\infty} \lambda_v \xi_v^2 \geq x\right)} = 1 + o(1), \tag{5.4}$$

uniformly over $x \in (0, o(N^\theta))$ where $\xi_1, \xi_2, \ldots$ are independent and identically distributed as $N(0, 1)$. Here θ is a constant that satisfies

$$\theta < \sup\left\{q \in [0, 1/3) : \sum\nolimits_{v > \lfloor N^{(1-3q)/5} \rfloor} \lambda_v = O(N^{-q})\right\},$$

when there exist infinitely many non-zero eigenvalues and $\theta = 1/3$ otherwise.

Remark 5.1. *Although we restrict our attention to the two-sample V-statistic with a second-order kernel $h^*(x_1, x_2; y_1, y_2)$, our result can be straightforwardly extended to higher-order kernels $h^*(x_1, \ldots, x_r; y_1, \ldots, y_r)$ for some $r \geq 3$. The key idea is to consider Hoeffding's decomposition of two-sample U-statistics (page 40 of Lee, 1990) and properly control the remainder terms (see, Drton et al., 2018, for one-sample case). Finally, using the relationship between U- and V-statistics (e.g. page 183 of Lee, 1990), one can derive the desired*

result for the V-statistic with a higher-order kernel. We do not pursue this direction here since the second-order kernel is enough for our application.

5.3.2 Gumbel limiting distribution

With the aid of Theorem 5.1, we are now ready to describe the limiting distribution of the proposed statistic under large K and large N situations. The main ingredient is Chen–Stein method for Poisson approximations (Arratia et al., 1989) that has been successfully applied to approximate the distribution of a maximum-type test statistic to a Gumbel distribution (e.g. Han et al., 2017; Drton et al., 2018). For sake of completeness, we state Theorem 1 of Arratia et al. (1989).

Lemma 5.1.1 (Theorem 1 of Arratia et al. (1989))**.** *Let $\mathcal{I}$ be an arbitrary index set and for $i \in \mathcal{I}$, let Y_i be a Bernoulli random variable with $p_i = \mathbb{P}(Y_i = 1) > 0$. For each $i \in \mathcal{I}$, consider a subset of $\mathcal{I}$ such that $B_i \subset \mathcal{I}$ with $i \in B_i$. Let us define $W = \sum_{i\in\mathcal{I}} Y_i$ and $\lambda = \mathbb{E}(W) = \sum_{i\in\mathcal{I}} p_i$. Let V be a Poisson random variable with mean λ. Then we have that*

$$\big|\mathbb{P}(W = 0) - \mathbb{P}(V = 0)\big| \leq \min\{1, \lambda^{-1}\}(b_1 + b_2 + b_3)$$

where

$$b_1 := \sum_{i\in\mathcal{I}} \sum_{j\in B_i} p_i p_j,\ b_2 := \sum_{i\in\mathcal{I}} \sum_{i\neq j\in B_i} \mathbb{E}(Y_i Y_j) \text{ and}$$
$$b_3 := \sum_{i\in\mathcal{I}} \mathbb{E}\Big|\mathbb{E}\Big[Y_i - p_i\Big| \sum_{j\in\mathcal{I}-B_i} Y_j\Big]\Big|.$$

Let us denote the two-sample MMD statistic between P_k and P_l by $\widehat{\mathcal{V}}^2_{kl}$, that is $\widehat{\mathcal{V}}^2_{kl} = \big\|n_k^{-1}\sum_{i=1}^{n_k}\psi(X_{i,k}) - n_l^{-1}\sum_{j=1}^{n_l}\psi(X_{j,l})\big\|^2_{\mathcal{H}}$. Assume the sample sizes are the same as $n := n_1 = \ldots = n_K$ for simplicity. Then based on the following key observation

$$\mathbb{P}\big(n\widehat{\mathcal{V}}^2_{h,\max}/2 \leq x\big) \ = \ \mathbb{P}\Big\{\sum\nolimits_{1\leq k<l\leq K} \mathbb{1}\big(n\widehat{\mathcal{V}}^2_{kl}/2 > x\big) = 0\Big\},$$

Lemma 5.1.1 can be applied in our context with $W = \sum_{1\leq k<l\leq K}\mathbb{1}\big(n\widehat{\mathcal{V}}^2_{kl}/2 > x\big)$ and $\lambda = \sum_{1\leq k<l\leq K}\mathbb{P}\big(n\widehat{\mathcal{V}}^2_{kl}/2 > x\big)$. Ultimately the proof boils down to showing that b_1, b_2, b_3 converge to zero under appropriate conditions. This has been established in Appendix D.1 and the result is summarized as follows.

Theorem 5.2 (Gumbel limit)**.** *Suppose that **(A1)** and **(A2)** are fulfilled. Consider a balanced sample case such that $n := n_1 = \ldots = n_K$. Let θ be a constant chosen as in Theorem 5.1 and assume that $\log K = o(n^\theta)$.*

Then under the null hypothesis $P_1 = \ldots = P_K$, for any $y \in \mathbb{R}$,

$$\begin{aligned} & \lim_{n,K\to\infty} \mathbb{P}\left(\frac{n}{2\lambda_1} \widehat{\mathcal{V}}^2_{h,\max} - 4\log K - (\mu_1 - 2)\log\log K \le y \right) \\ = \;& \exp\left\{ - \frac{2^{\mu_1/2-2}\kappa}{\Gamma(\mu_1/2)} \exp\left(-\frac{y}{2}\right) \right\}, \end{aligned}$$

where $\kappa = \prod_{v=\mu_1+1}^{\infty}(1 - \lambda_v/\lambda_1)^{-1/2}$ and μ_1 is the multiplicity of the largest eigenvalue among the sequence $\{\lambda_v\}_{v=1}^{\infty}$.

Remark 5.2. *From Theorem 5.2, it is clear that we need to know or at least estimate a potentially infinite number of parameters $\{\lambda_v\}_{v=1}^{\infty}$ in order to implement the asymptotic test. Even if one has access to these eigenvalues, the asymptotic test might suffer from slow convergence. This means that the test can be too liberal or too conservative in finite sample size situations.*

Remark 5.3. *When the sample sizes are unbalanced, the limiting distribution of $\widehat{\mathcal{V}}^2_{h,max}$ may not have an explicit expression as in Theorem 5.2. In particular, it depends on the limit values of $n_k/(n_k + n_l)$ for $1 \le k < l \le K$. To avoid this complication, we simply focus on the case of equal sample sizes and present the explicit formula for the limiting distribution. Nevertheless, if we instead use the weighted K-sample statistic:*

$$\max_{1\le k<l\le K} \left(\frac{n_k n_l}{n_k + n_l} \widehat{\mathcal{V}}^2_{kl} \right),$$

we may obtain the same Gumbel limiting distribution as in Theorem 5.2 for general sample sizes.

5.3.3 Examples

In general, it is challenging to find closed-form expressions for $\{\lambda_v\}_{v=1}^{\infty}$ and $\{\varphi_v(\cdot)\}_{v=1}^{\infty}$ as they depend on the kernel as well as the underlying distribution. We end this section with two simple examples for which $\{\lambda_v\}_{v=1}^{\infty}$ and $\{\varphi_v(\cdot)\}_{v=1}^{\infty}$ are explicit. Based on these, we illustrate Theorem 5.2.

- **Linear kernel:** Suppose that $\{X_{1,1}, \ldots, X_{n,1}, \ldots X_{1,K}, \ldots, X_{n,K}\}$ are independent and identically distributed as a multivariate normal distribution with mean zero and covariance matrix Σ. Suppose further that Σ is a diagonal matrix whose diagonal entries are $\lambda_1 = \ldots = \lambda_{\mu_1} > \lambda_{\mu_1+1} \ge \ldots \ge \lambda_d > 0$ for some $\mu_1 \ge 1$. Let us consider the linear kernel given as $h(x_1, x_2) = x_1^\top x_2$. Then it is straightforward to see that the centered kernel in (5.2) has the eigenfunction decomposition as

$$\overline{h}(x_1, x_2) = \sum_{v=1}^{d} \lambda_v \varphi_v(x_1)\varphi_v(x_2) = \sum_{v=1}^{d} \lambda_v \big(x_1^{(v)}/\sqrt{\lambda_v}\big)\big(x_2^{(v)}/\sqrt{\lambda_v}\big)$$

where $x_1^{(v)}$ is the vth component of x_1. Under the given setting, $\{\varphi_1(X_{1,1}), \ldots, \varphi_d(X_{1,1})\}$ are independent and identically distributed as $N(0,1)$. It can be shown that the conditions in Theorem 5.2 are satisfied with $\theta = 1/3$ under the Gaussian assumption. Thus the resulting test statistic converges to a Gumbel distribution as in Theorem 5.2.

- **Chi-square kernel:** Suppose that $\{X_{1,1}, \ldots, X_{n,1}, \ldots X_{1,K}, \ldots, X_{n,K}\}$ are independent and identically distributed on a discrete domain $\{1, \ldots, m\}$ with fixed m. Let $p_v > 0$ be the probability of observing the value v among $\{1, \ldots, m\}$ and consequently $\sum_{v=1}^m p_v = 1$. Consider the chi-square kernel defined as $h(x_1, x_2) = \sum_{v=1}^m p_v^{-1} \mathbb{1}(x_1 = v)\mathbb{1}(x_2 = v)$. Let A be a $(m-1) \times (m-1)$ matrix whose (v_1, v_2) entry is $a_{v_1,v_2} = p_{v_1}^{-1} + p_m^{-1}$ if $v_1 = v_2$ and $a_{v_1,v_2} = p_m^{-1}$ otherwise. Let us define the eigenfunction $\varphi_v(x)$ to be the vth row of $A^{1/2}\{\mathbb{1}(x = 1) - p_1, \ldots, \mathbb{1}(x = m-1) - p_{m-1}\}^\top$ for $v = 1, \ldots, m-1$. Then, following the calculation in Theorem 14.3.1 of Lehmann and Romano (2006),

$$\overline{h}(x_1, x_2) = \sum_{v=1}^{m-1} \lambda_v \varphi_v(x_1)\varphi_v(x_2) = \sum_{v=1}^{m} \frac{\{\mathbb{1}(x_1 = v) - p_v\}\{\mathbb{1}(x_2 = v) - p_v\}}{p_v},$$

where $\lambda_1 = \ldots = \lambda_{m-1} = 1$ and $\lambda_v = 0$ for $v \geq m$ and the eigenfunctions are bounded. Thus the conditions in Theorem 5.2 are satisfied with $\theta = 1/3$ and the resulting test statistic converges to a Gumbel distribution.

5.4 Permutation Approach

So far we have investigated the limiting null distribution of the proposed test statistic and demonstrated the difficulty of implementing the resulting asymptotic test. To address the issue, we take an alternative approach based on permutations that does not require prior knowledge on unknown parameters. The key advantage of the permutation approach is that it yields a valid level α test (or a size α test via randomization) for any finite sample size and for any number of distributions. This attractive property is true for any type of underlying distributions, provided that $\{Z_1, \ldots, Z_N\}$ are exchangeable under H_0. In the following, we briefly describe the original and randomized permutation procedures. The randomized procedure has a computational advantage over the original procedure by considering a random subset of all permutations.

- **Permutation approach**: Let $\mathcal{B}_N$ be the collection of all possible permutations of $\{1, \ldots, N\}$. For $\boldsymbol{b} = (b_1, \ldots, b_N) \in \mathcal{B}_N$, we denote by $\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})}$ the test statistic computed based on the permuted dataset $\{Z_{b_1}, \ldots, Z_{b_N}\}$. We then clearly have $\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b}_0)} = \widehat{\mathcal{V}}_{h,\max}$ for $\boldsymbol{b}_0 = (1, \ldots, N)$. The permutation p-value is calculated by

$$p_{\text{perm}} = \frac{1}{N!} \sum_{\boldsymbol{b} \in \mathcal{B}_N} \mathbb{1}\left(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \geq \widehat{\mathcal{V}}_{h,\max}\right). \tag{5.5}$$

It is well-known that $\mathbb{P}(p_{\text{perm}} \leq t) \leq t$ for any $0 \leq t \leq 1$ under H_0 (e.g. Chapter 15 of Lehmann and Romano, 2006). Consequently $\mathbb{1}(p_{\text{perm}} \leq \alpha)$ is a valid level α test.

- **Randomized version**: For large N, it would be beneficial to consider a subset of $\mathcal{B}_N$ and compute the approximated permutation p-value. Suppose that $\boldsymbol{b}'_1, \ldots, \boldsymbol{b}'_M$ are sampled uniformly from $\mathcal{B}_N$ with replacement. We then define a Monte-Carlo version of the permutation p-value by

$$p_{\text{MC}} = \frac{1}{M+1}\left\{1 + \sum_{i=1}^{M} \mathbb{1}\left(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b}'_i)} \geq \widehat{\mathcal{V}}_{h,\max}\right)\right\}. \tag{5.6}$$

  It can be shown that $\mathbb{P}(p_{\text{MC}} \leq t) \leq t$ for any $0 \leq t \leq 1$ under H_0 (e.g. Chapter 15 of Lehmann and Romano, 2006). Hence $\mathbb{1}(p_{\text{MC}} \leq \alpha)$ is a valid level α test as well.

Having motivated the permutation approach, we next analyze uniform consistency as well as minimax optimality of the resulting permutation test against sparse alternatives in Section 5.6, building on concentration inequalities developed in the following section.

5.5 Concentration inequalities under permutations

This section develops a concentration inequality for the permuted MMD statistic with an exponential tail bound. The result established here is especially useful for studying the type II error (or the power) of the proposed permutation test in Section 5.6. Our result can also be valuable in addressing the computational issue of the permutation test. The permutation approach suffers from high computational cost as the number of all possible permutations increases very quickly with the sample size. As a result, it is common in practice to use Monte-Carlo sampling of random permutations to approximate the p-value of a permutation test. However, in some application areas such as genetic where extremely small p-values are of interest, the Monte-Carlo approach still requires heavy computations (Knijnenburg et al., 2009; He et al., 2019). Our concentration inequality has an exponential tail bound with completely known quantities. Based on this, one can find a sharp upper bound for the permutation p-value (or the permutation critical value) without any computational cost for permutations. We discuss this direction in more detail in Remark 5.5.

5.5.1 Bobkov's inequality

Before we state the main result of this section, we introduce Bobkov's inequality (Bobkov, 2004), which is the key ingredient of our proof. To state his result, we need to prepare some notation in advance. Consider a discrete cube given by

$$\mathcal{G}_{N,m} = \{\boldsymbol{w} = (w_1, \ldots, w_N) \in \{0,1\}^N : w_1 + \ldots w_N = m\}.$$

Note that for each $\boldsymbol{w} \in \mathcal{G}_{N,m}$, there are exactly $m(N-m)$ neighbors $\{s_{ij}\boldsymbol{w}\}_{i\in I(\boldsymbol{w}),j\in J(\boldsymbol{w})}$ where $I(\boldsymbol{w}) = \{i \leq N : w_i = 1\}$ and $J(\boldsymbol{w}) = \{j \leq N : w_j = 0\}$ such that $(s_{ij}\boldsymbol{w})_r = w_r$ for $r \neq i, j$ and $(s_{ij}\boldsymbol{w})_i = w_j, (s_{ij}\boldsymbol{w})_j = w_i$. Now for a function f defined on $\mathcal{G}_{N,m}$, the Euclidean length of discrete gradient $\nabla f(\boldsymbol{w})$ is given as

$$|\nabla f(\boldsymbol{w})|^2 = \sum_{i\in I(\boldsymbol{w})} \sum_{j\in J(\boldsymbol{w})} |f(\boldsymbol{w}) - f(s_{ij}\boldsymbol{w})|^2.$$

For more details, we refer to Bobkov (2004). Then Bobkov's inequality is stated as follows:

Lemma 5.2.1 (Theorem 2.1 of Bobkov (2004))**.** *For every real-valued function f on $\mathcal{G}_{N,m}$ and $|\nabla f(\boldsymbol{w})| \leq \Sigma$ for all $\boldsymbol{w}$,*

$$\mathbb{P}_{\boldsymbol{w}}[f(\boldsymbol{w}) - \mathbb{E}_{\boldsymbol{w}}\{f(\boldsymbol{w})\} \geq t] \leq \exp\{-(N+2)t^2/(4\Sigma^2)\},$$

where $\mathbb{P}_{\boldsymbol{w}}(\cdot)$ represents a counting probability measure on $\mathcal{G}_{N,m}$ and $\mathbb{E}_{\boldsymbol{w}}(\cdot)$ is the expectation associated with $\mathbb{P}_{\boldsymbol{w}}(\cdot)$.

5.5.2 Two-Sample Case

We first focus on the two-sample case. When $K = 2$, it is clear that the proposed test statistic becomes the V-statistic in Gretton et al. (2012) and

$$\widehat{\mathcal{V}}_{h,\max} = \frac{N}{n_2}\left\|\frac{1}{n_1}\sum_{i_1=1}^{n_1}\psi(X_{i,1}) - \frac{1}{N}\sum_{j=1}^{N}\psi(Z_j)\right\|_{\mathcal{H}} = \frac{N}{n_1 n_2}\left\|\sum_{i_1=1}^{n_1}\overline{\psi}(Z_i)\right\|_{\mathcal{H}}, \tag{5.7}$$

where $\overline{\psi}(Z_{i_1}) = \psi(Z_{i_1}) - \frac{1}{N}\sum_{j=1}^{N}\psi(Z_j)$. Recall that $\boldsymbol{b}$ is a N-dimensional random vector uniformly distributed over $\mathcal{B}_N$ in the permutation procedure. As before in Section 5.4, we denote the test statistic based on the permuted dataset $\{Z_{b_1}, \ldots, Z_{b_N}\}$ by

$$\widehat{\mathcal{V}}^{(\boldsymbol{b})}_{h,\max} := \frac{N}{n_1 n_2}\left\|\sum_{i_1=1}^{n_1}\overline{\psi}(Z_{b_{i_1}})\right\|_{\mathcal{H}}.$$

We also denote the probability law under permutations (conditional on $Z_1, \ldots, Z_N$) by $\mathbb{P}_{\boldsymbol{b}}(\cdot)$ and the expectation associated with $\mathbb{P}_{\boldsymbol{b}}(\cdot)$ by $\mathbb{E}_{\boldsymbol{b}}(\cdot)$.

It should be stressed that in the two-sample case, there exists $\boldsymbol{w} \in \mathcal{G}_{N,n_1}$ corresponding to each $\boldsymbol{b} \in \mathcal{B}_N$ such that

$$\widehat{\mathcal{V}}^{(\boldsymbol{b})}_{h,\max} = \widehat{\mathcal{V}}^{[\boldsymbol{w}]}_{h,\max} := \frac{N}{n_1 n_2}\left\|\sum_{i=1}^{N} w_i\overline{\psi}(Z_i)\right\|_{\mathcal{H}}.$$

More importantly, both $\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})}$ and $\widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]}$ have the same probability law when $\boldsymbol{b}$ and $\boldsymbol{w}$ are uniformly distributed over $\mathcal{B}_N$ and $\mathcal{G}_{N,n_1}$, respectively. In other words, we have

$$\mathbb{P}_{\boldsymbol{b}}\big\{\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} - \mathbb{E}_{\boldsymbol{b}}(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})}) \geq t\big\} = \mathbb{P}_{\boldsymbol{w}}\big\{\widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]} - \mathbb{E}_{\boldsymbol{w}}(\widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]}) \geq t\big\} \quad \text{for all } t \in \mathbb{R}.$$

This key observation allows us to apply Bobkov's inequality to obtain a concentration inequality for the permuted test statistic in the following theorem.

Theorem 5.3 (Concentration inequality for two-sample statistic)**.** *For $K = 2$, let $\mathbb{P}_{\boldsymbol{b}}$ be the uniform probability measure over permutations conditional on $\{Z_1, \ldots, Z_N\}$. Let us write $\gamma_{1,2} = n_1 n_2/(n_1+n_2)^2$. Further denote $\widetilde{h}(Z_i, Z_j) = h(Z_i, Z_i) + h(Z_j, Z_j) - 2h(Z_i, Z_j) \geq 0$ and*

$$\widehat{\sigma}^2 = \frac{1}{N(N-1)} \sum_{i \neq j = 1}^{N} \widetilde{h}(Z_i, Z_j). \tag{5.8}$$

Then for all $t > 0$, we have

$$\mathbb{P}_{\boldsymbol{b}}\left(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \geq t + \sqrt{\frac{\widehat{\sigma}^2}{2N\gamma_{1,2}}}\right) \leq \exp\left(-\frac{N\gamma_{1,2}^2 t^2}{2\widehat{\sigma}^2}\right). \tag{5.9}$$

Proof. From the previous discussion, it suffices to investigate a concentration inequality for $f(\boldsymbol{w}) := \widehat{\mathcal{V}}_{h,\max}^{[\boldsymbol{w}]}$, which is uniformly distributed on $\mathcal{G}_{N,n_1}$. Since Bobkov's inequality holds for $f(\boldsymbol{w})$, all we need to do is to find meaningful bounds of the expected value of $f(\boldsymbol{w})$ and the Euclidean length of $\nabla f(\boldsymbol{w})$. We first bound the expected value of $f(\boldsymbol{w})$. Using the feature map representation of kernel h, it is straightforward to see that

$$\sum_{i=1}^{N} \|\overline{\psi}(Z_i)\|_{\mathcal{H}}^2 = -\sum_{i \neq j = 1}^{N} \big\langle \overline{\psi}(Z_i), \overline{\psi}(Z_j) \big\rangle_{\mathcal{H}} = \frac{1}{2N} \sum_{i \neq j = 1}^{N} \widetilde{h}(Z_i, Z_j). \tag{5.10}$$

Then using Jensen's inequality together with the above identities,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}}\left[\left\| \sum_{i=1}^{N} w_i \overline{\psi}(Z_i) \right\|_{\mathcal{H}}\right] &\leq \sqrt{\mathbb{E}_{\boldsymbol{w}}\left[\sum_{i=1}^{N} w_i^2 \left\|\overline{\psi}(Z_i)\right\|_{\mathcal{H}}^2 + \sum_{i \neq j = 1}^{N} w_i w_j \big\langle \overline{\psi}(Z_i), \overline{\psi}(Z_i) \big\rangle_{\mathcal{H}}\right]} \\
&= \sqrt{\frac{n_1}{N} \sum_{i=1}^{N} \left\|\overline{\psi}(Z_i)\right\|_{\mathcal{H}}^2 + \frac{n_1(n_1-1)}{N(N-1)} \sum_{i \neq j = 1}^{N} \big\langle \overline{\psi}(Z_i), \overline{\psi}(Z_j) \big\rangle_{\mathcal{H}}} \\
&= \sqrt{\frac{n_1 n_2}{2N^2(N-1)} \sum_{i \neq j = 1}^{N} \widetilde{h}(Z_i, Z_j)}.
\end{aligned}$$

By multiplying the scaling factor $N/(n_1 n_2)$ on both sides, we have $\mathbb{E}_{\boldsymbol{w}}[f(\boldsymbol{w})] \leq \sqrt{\widehat{\sigma}^2/(2N\gamma_{1,2})}$.

Next we bound $|\nabla f(\boldsymbol{w})|$. Recall the definition of $s_{ij}w$ in Section 5.5.1. Using the triangle inequality, we see that

$$\left| \frac{N}{n_1 n_2} \left\| \sum_{l=1}^{N} w_l \overline{\psi}(Z_l) \right\|_{\mathcal{H}} - \frac{N}{n_1 n_2} \left\| \sum_{l=1}^{N} (s_{ij}w)_l \overline{\psi}(Z_l) \right\|_{\mathcal{H}} \right| \leq \frac{N}{n_1 n_2} \left\| \overline{\psi}(Z_i) - \overline{\psi}(Z_j) \right\|_{\mathcal{H}}.$$

Based on this observation, one can find Σ, which is independent of $\boldsymbol{w}$, as

$$|\nabla f(\boldsymbol{w})|^2 \leq \Sigma^2 := \frac{N^2}{n_1^2 n_2^2} \sum_{1 \leq i < j \leq N} \left\| \overline{\psi}(Z_i) - \overline{\psi}(Z_j) \right\|_{\mathcal{H}}^2 = \frac{N^2}{2n_1^2 n_2^2} \sum_{i \neq j = 1}^{N} \widetilde{h}(Z_i, Z_j),$$

where the last equality uses the identities in (5.10). Now apply Bobkov's inequality with the above pieces to obtain the desired result. □

Remark 5.4. *Before we move on, we make several comments on Theorem 5.3.*

(a) The tail of the given concentration inequality relies solely on the variance term of the kernel. This is in sharp contrast to Hoeffding or Bernstein-type inequalities (e.g. Boucheron et al., 2013) that usually depend on the (possibly unknown) range of random variables.

(b) The given concentration inequality requires no assumption on random variables such as boundedness or more generally sub-Gaussianity. Furthermore it only depends on known and easily computable quantities in practice.

(c) For $0 < \alpha < 1$, consider a test function $\phi_2 : \{Z_1, \ldots, Z_N\} \mapsto \{0, 1\}$ such that

$$\phi_2 = I\left\{ \widehat{\mathcal{V}}_{h,max} \geq \sqrt{\frac{2\widehat{\sigma}^2}{N\gamma_{1,2}^2} \log\left(\frac{1}{\alpha}\right)} + \sqrt{\frac{\widehat{\sigma}^2}{2N\gamma_{1,2}}} \right\}.$$

As a corollary of Theorem 5.3, it can be seen that ϕ_2 is a valid level α test whenever $\{Z_1, \ldots, Z_N\}$ are exchangeable.

(d) We stress that our test statistic is a degenerate two-sample V-statistic. Therefore, the previous studies on concentration inequalities for the permuted simple sum (e.g. Chatterjee, 2007; Adamczak et al., 2016; Albert, 2018) cannot be applied in our context.

5.5.3 Numerical Illustrations

We illustrate the usefulness of Theorem 5.3 via simulations. First of all, we can use Theorem 5.3 to compute an upper bound for the original permutation p-value. In detail, suppose that $n_1 = n_2$ with $N = n_1 + n_2$ for

simplicity. Then it is straightforward to see that the permutation p-value is less than or equal to

$$p_{\text{Bobkov}} := \begin{cases} \exp\left\{ -\frac{N}{32\widehat{\sigma}^2}\left(\widehat{\mathcal{V}}_{h,\max} - \sqrt{\frac{2\widehat{\sigma}^2}{N}}\right)^2\right\}, & \text{if } \widehat{\mathcal{V}}_{h,\max} \geq \sqrt{\frac{2\widehat{\sigma}^2}{N}} \\ 1, & \text{else.} \end{cases}$$

By the nature of the permutation test, p_{Bobkov} is a valid p-value in any finite sample size, in a sense that $\mathbb{P}(p_{\text{Bobkov}} \leq \alpha) \leq \alpha$ under H_0. Another way of obtaining a finite-sample valid p-value is to use an *unconditional* concentration inequality. For example, Gretton et al. (2012) employ McDiarmid's inequality (McDiarmid, 1989) to have an concentration inequality for the MMD V-statistic with a bounded kernel. Based on Theorem 7 of Gretton et al. (2012) under the bounded kernel assumption $0 \leq h(x, y) \leq B$, another valid p-value can be obtained as

$$p_{\text{McDiarmid}} := \begin{cases} \exp\left\{ -\frac{N}{8B}\left(\widehat{\mathcal{V}}_{h,\max} - \sqrt{\frac{32B}{N}}\right)^2\right\}, & \text{if } \widehat{\mathcal{V}}_{h,\max} \geq \sqrt{\frac{32B}{N}} \\ 1, & \text{else.} \end{cases}$$

Both approaches provide exponentially decaying p-values in sample size but we should emphasize that p_{Bobkov} does not require any moment conditions on the kernel. Even if the kernel is bounded, p_{Bobkov} would be preferred to $p_{\text{McDiarmid}}$ when $\widehat{\sigma}^2$ is much smaller than B. This point is illustrated under the following set-up.

Set-up. We consider two kernels: 1) energy distance kernel $h(x, y) = (\|x\|_2 + \|y\|_2 - \|x - y\|_2)/2$ and 2) linear kernel $h(x, y) = x^\top y$. Although these kernels are unbounded in general, they are bounded when the underlying distributions have compact support. For this purpose, we consider two truncated normal distributions with the different location parameters $\mu_1 = 1$ and $\mu_2 = -1$ and the same scale parameter $\sigma^2 = 1$. We let both distributions have the same support as $[-5, 5]$ so that we can calculate the bound B for each kernel. For each sample size N among $\{100, 200, \ldots, 900, 1000\}$, the experiments were repeated 200 times to estimate the expected values of the p-values.

Results. In Figure 5.1, we present the simulation results of the comparison between p_{Bobkov} and $p_{\text{McDiarmid}}$ under the described scenario. The p-values are displayed in log-scale for better visual comparison. Under the given setting, we observe that $\widehat{\sigma}^2$ is much smaller than B for both kernels, which in turns leads to a smaller value of p_{Bobkov} compared to $p_{\text{McDiarmid}}$. More specifically, we observe 1) $\widehat{\sigma}^2 \approx 1.61$ on average and $B = 10$ for the energy distance kernel and 2) $\widehat{\sigma}^2 \approx 4.01$ on average and $B = 100$ for the linear kernel. It is worth noting that the benefit of using p_{Bobkov} becomes more evident for unbounded random variables for which $p_{\text{McDiarmid}}$ is not even applicable.

Remark 5.5. *The test based on p_{Bobkov} may not be recommended when the sample size is small and the significance level α is of moderate size (e.g. $\alpha = 0.05$). In this case, the permutation test via Monte-Carlo*

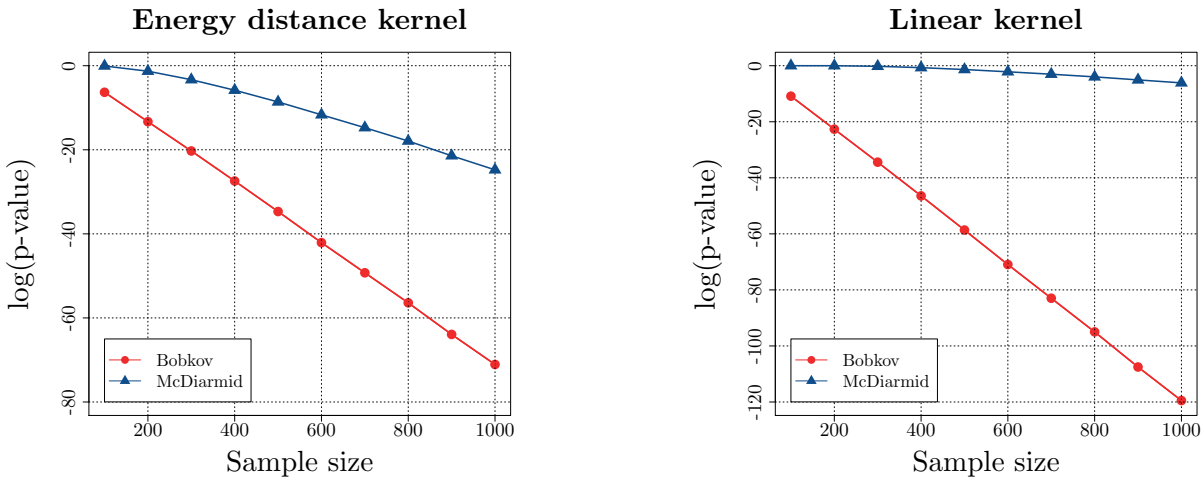

Figure 5.1: Comparisons between Bobkov's inequality and McDiarmid inequality in their application to p-value evaluation. In both energy distance kernel and linear kernel, Bobkov's inequality returns significantly smaller p-values than McDiarmid inequality. See Section 5.5.3 for details.

simulations would be more satisfactory. However, when the sample size is large and the significance level is very small (e.g. $\alpha = 10^{-100}$*), the Monte-Carlo approach would be computationally infeasible, requiring at least* α^{-1} *random permutations in order to reject* H_0*. In this large-sample and small* α *situation, the approach based on* p_{Bobkov} *would be practically valuable, which does not require any computational cost on permutations.*

Remark 5.6. *While we focused on the case where* $\widehat{\sigma}^2 \ll B$ *to highlight the advantage of* p_{Bobkov}*, it is definitely possible to observe that* $p_{McDiarmid}$ *is smaller than* p_{Bobkov}*, especially when* B *is comparable to or smaller than* $\widehat{\sigma}^2$*.*

5.5.4 K-Sample Case

Next we give a general result for arbitrary $K \geq 2$. Unfortunately, we cannot directly apply Bobkov's inequality when $K > 2$ since the inequality holds only for a function $f(\boldsymbol{w})$ defined on a binary discrete cube. Our strategy to overcome this problem is to first apply Bobkov's inequality to each pairwise MMD test statistic and then aggregate the results via the union bound. To start, we introduce $\widehat{\sigma}_K^2$ in Algorithm 3 that generalizes $\widehat{\sigma}^2$ to the K-sample case.

It can be seen that $\widehat{\sigma}_K^2$ is the same as $\widehat{\sigma}^2$ in (5.8) when $K = 2$ and can be computed in quadratic time for large K. Using $\widehat{\sigma}_K^2$, we extend Theorem 5.3 as follows.

Theorem 5.4 (Concentration inequality for K-sample statistic)**.** *For* $K \geq 2$*, let* $\mathbb{P}_{\boldsymbol{b}}$ *be the uniform probability measure over permutations conditional on* $\{Z_1, \ldots, Z_N\}$*. For distinct* $k, l \in \{1, \ldots, K\}$*, let* $\gamma_{k,l} = n_k n_l/(n_k +$

Algorithm 3: Calculation of $\widehat{\sigma}_K^2$

Require: the pooled samples $\{Z_1, \ldots, Z_N\}$, the number of samples $n_1, \ldots, n_K$.

(1) Calculate $\widetilde{h}(Z_i, Z_j)$ for $1 \le i \ne j \le N$.

(2) Sort and denote the previous outputs by $\widetilde{h}_{[1]} \ge \ldots \ge \widetilde{h}_{[N(N-1)]}$.

(3) Compute $\widehat{\sigma}_K^2 := \max_{1 \le k < l \le K} \overline{\sigma}_{kl}^2$ where $\overline{\sigma}_{kl}^2$ is the sample average of $\widetilde{h}_{[1]}, \widetilde{h}_{[2]}, \ldots, \widetilde{h}_{[(n_k+n_l)(n_k+n_l-1)]}$.

(4) Return $\widehat{\sigma}_K^2$.

$n_l)^2$ *and consider* $\widehat{\sigma}_K^2$ *in Algorithm 3. Then for any* $t \ge 0$,

$$\mathbb{P}_{\boldsymbol{b}}\left\{\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \ge t + \max_{1 \le k < l \le K} \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right\} \le \binom{K}{2} \exp\left\{-\min_{1 \le k < l \le K} \frac{(n_k+n_l)\gamma_{k,l}^2 t^2}{2\widehat{\sigma}_K^2}\right\}. \tag{5.11}$$

Proof. For a given permutation $\boldsymbol{b} \in \mathcal{B}_N$, let us denote

$$\widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})} = \left\| \frac{1}{n_k} \sum_{i=1}^{n_k} \psi(Z_{b_{m_{k-1}+i}}) - \frac{1}{n_l} \sum_{j=1}^{n_l} \psi(Z_{b_{m_{l-1}+j}}) \right\|_{\mathcal{H}},$$

where $m_{l-1} = \sum_{k=1}^{l-1} n_k$ and $m_0 = 0$ so that $\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} = \max_{1 \le k < l \le K} \widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})}$. Based on the triangle inequality and the union bound, observe that

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{b}}\left\{\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \ge t + \max_{1 \le k < l \le K} \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right\} &\le \mathbb{P}_{\boldsymbol{b}}\left[\max_{1 \le k < l \le K}\left\{\widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})} - \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right\} \ge t\right] \\
&\le \sum_{1 \le k < l \le K} \mathbb{P}_{\boldsymbol{b}}\left\{\widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})} \ge t + \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right\}.
\end{aligned} \tag{5.12}$$

Let $\widetilde{Z} = \{\widetilde{Z}_1, \ldots, \widetilde{Z}_{n_k+n_l}\}$ be the $n_k + n_l$ samples uniformly drawn from $\{Z_1, \ldots, Z_N\}$ without replacement. Write

$$\widehat{\mathcal{V}}_{kl}^{[\boldsymbol{w}]} = \frac{n_k+n_l}{n_k n_l} \left\| \sum_{i_1=1}^{n_k+n_l} w_{i_1} \left\{\psi(\widetilde{Z}_{i_1}) - \frac{1}{n_k+n_l} \sum_{i_2=1}^{n_k+n_l} \psi(\widetilde{Z}_{i_2})\right\} \right\|_{\mathcal{H}},$$

where $\boldsymbol{w} = \{w_1, \ldots, w_{n_k+n_l}\}$ is a set of Bernoulli random variables uniformly distributed on $\mathcal{G}_{n_k+n_l,n_k}$ as before. Then by the law of total expectation and a slight modification of the proof of Theorem 5.3, it can be seen that

$$\mathbb{P}_{\boldsymbol{b}}\left(\widehat{\mathcal{V}}_{kl}^{(\boldsymbol{b})} \ge t + \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right) = \mathbb{E}_{\widetilde{Z}}\left[\mathbb{P}_{\boldsymbol{w}}\left\{\widehat{\mathcal{V}}_{kl}^{[\boldsymbol{w}]} \ge t + \sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}} \;\middle|\; \widetilde{Z}\right\}\right]$$

$$\leq \mathbb{E}_{\widetilde{Z}}\left[\exp\left\{-\frac{(n_k+n_l)\gamma_{k,l}^2 t^2}{2\widehat{\sigma}_K^2}\right\}\right]$$
$$= \exp\left\{-\frac{(n_k+n_l)\gamma_{k,l}^2 t^2}{2\widehat{\sigma}_K^2}\right\},$$

where the last equality follows since $\widehat{\sigma}_K^2$ is invariant to the choice of $\widetilde{Z}$. By putting this result into the right-hand side of (5.12), the proof is complete. □

Remark 5.7. *We provide some comments on Theorem 5.4.*

(a) When $K = 2$, the concentration inequality given in (5.11) recovers the one in (5.9).

(b) One can replace $\widehat{\sigma}_K^2$ with $\max_{1\leq i<j\leq N}\widetilde{h}(Z_i, Z_j)$ in (5.11), which takes less time to compute, but at the expense of the loss of the tightness. Note, however, that the bound with $\max_{1\leq i<j\leq N}\widetilde{h}(Z_i, Z_j)$ is tight enough to prove minimax rate optimality of the proposed test. See the proof of Theorem 5.5 for details.

(c) As before in the two-sample case, the proposed K-sample concentration inequality is valid without any moment condition and it depends solely on known and easily computable quantities.

(d) Consider a test function $\phi_K : \{Z_1, \ldots, Z_N\} \mapsto \{0, 1\}$ such that

$$\phi_K = I\left[\widehat{\mathcal{V}}_{h,max} \geq \max_{1\leq k<l\leq K}\sqrt{\left\{\frac{2\widehat{\sigma}_K^2}{(n_k+n_l)\gamma_{k,l}^2}\right\}\log\left\{\frac{\binom{K}{2}}{\alpha}\right\}} + \max_{1\leq k<l\leq K}\sqrt{\frac{\widehat{\sigma}_K^2}{2(n_k+n_l)\gamma_{k,l}}}\right].$$

As a corollary of Theorem 5.4, it can be seen that ϕ_K is a valid level α test whenever $\{Z_1, \ldots, Z_N\}$ are exchangeable under H_0.

5.6 Power Analysis

In this section, we study the power of the permutation test based on the proposed test statistic and prove its minimax rate optimality against certain sparse alternatives. Throughout this section, we need the following assumptions:

(B1). Assume that kernel h is uniformly bounded by $0 \leq h(x, y) \leq B$ for all $x, y \in \mathcal{X}$.

(B2). There exists a fixed constant $c > 0$ such that $n_{\max}/n_{\min} \leq c$ for any sample sizes where $n_{\max}$ and $n_{\min}$ are the maximum and the minimum of $\{n_1, \ldots, n_K\}$ respectively.

Note that the assumption **(B1)** is satisfied by some widely used kernels e.g. Gaussian and Laplace kernels. It can also be satisfied by many other kernels when the underlying distributions have compact support. The

second assumption **(B2)** states that $n_1, \ldots, n_K$ are well-balanced. This assumption, for example, holds for the equal sample sizes with $c = 1$.

5.6.1 Power of the permutation test

Let $\mathcal{P}$ be the set of all distributions on $(\mathcal{X}, \mathcal{B})$. We characterize the difference between the null and the alternative in terms of $\max_{1\le k<l\le K} \mathcal{V}_h(P_k, P_l)$, which is the population counterpart of the proposed test statistic $\widehat{\mathcal{V}}_{h,\max}$. In particular, for a given positive sequence ϵ_N and kernel h, let us define a class of alternatives:

$$\mathcal{F}_h(\epsilon_N) = \big\{(P_1, \ldots, P_K) \in \mathcal{P} : \max_{1\le k<l\le K} \mathcal{V}_{kl} \ge \epsilon_N\big\}, \tag{5.13}$$

where $\mathcal{V}_{kl} = \mathcal{V}_h(P_k, P_l)$ for simplicity. We call the collection of alternatives in $\mathcal{F}_h(\epsilon_N)$ as the sparse alternatives, in a sense that only a few of $\{\mathcal{V}_{kl}\}_{1\le k<l\le K}$ are required to be greater than ϵ_N while the rest of them can be zero. Such sparse alternatives have been considered by many authors including Cai et al. (2013), Cai et al. (2014) and Han et al. (2017) in different contexts. The main goal of this subsection is to characterize the conditions under which the permutation test can be uniformly powerful over $\mathcal{F}_h(\epsilon_N)$. More specifically, we show that as long as the lower bound ϵ_N is sufficiently larger than

$$r_N^\star := \sqrt{\frac{\log K}{n_{\min}}},$$

then the proposed permutation test is uniformly consistent. Furthermore, in Section 5.6.2, we prove that this rate cannot be improved from a minimax perspective under some mild conditions on kernel h. In other words, the proposed test is minimax rate optimal against the sparse alternatives with the minimax rate $r_N^\star$.

We start by providing one lemma, which states that $\max_{1\le k<l\le K} |\widehat{\mathcal{V}}_{kl} - \mathcal{V}_{kl}|$ is bounded by $C\sqrt{\log K / n_{\min}}$ for some constant C with high probability.

Lemma 5.4.1. *Suppose that* **(B1)** *holds and recall that* $\widehat{\mathcal{V}}_{kl} = \big\|n_k^{-1}\sum_{i_1=1}^{n_k}\psi(X_{i_1,k}) - n_l^{-1}\sum_{i_2=1}^{n_l}\psi(X_{i_2,l})\big\|_{\mathcal{H}}$. *Then with probability at least* $1-\beta$ *where* $0 < \beta < 1$, *we have*

$$\max_{1\le k<l\le K} \big|\widehat{\mathcal{V}}_{kl} - \mathcal{V}_{kl}\big| \le 4\sqrt{\frac{B}{n_{\min}}} + 2\sqrt{\frac{B}{n_{\min}}\log\left\{\frac{2}{\beta}\binom{K}{2}\right\}}.$$

Proof. Using Theorem 7 of Gretton et al. (2012), one can obtain

$$\mathbb{P}\left(\big|\widehat{\mathcal{V}}_{kl} - \mathcal{V}_{kl}\big| \ge 2\sqrt{n_k^{-1}B} + 2\sqrt{n_l^{-1}B} + t\right) \le 2\exp\left\{-\frac{(n_k+n_l)\gamma_{k,l}t^2}{2B}\right\}.$$

Then the result follows by applying the union bound as in Theorem 5.4 and the following inequality

$$\min_{1\leq k<l\leq K}(n_k+n_l)\gamma_{k,l}\geq\frac{n_{\min}}{2}. \qquad \square$$

By building on Theorem 5.4 and Lemma 5.4.1, we prove the uniform consistency of the permutation test against $\mathcal{F}_h(\epsilon_N)$ when ϵ_N is much larger than $r_N^\star$. We provide the proof in Appendix D.1.

Theorem 5.5 (Uniform consistency of the original permutation test)**.** *Assume that* **(B1)** *and* **(B2)** *are fulfilled. Denote the permutation test function by* $\phi_{K,\mathrm{perm}}=\mathbb{1}(p_{\mathrm{perm}}\leq\alpha)$ *where* p_{perm} *is given in (5.5). Then under* H_1*,*

$$\limsup_{n_{\min}\to\infty}\sup_{(P_1,\ldots,P_K)\in\mathcal{F}_h(b_N r_N^\star)}\mathbb{P}\left(\phi_{K,\mathrm{perm}}=0\right)=0,$$

where b_N *is an arbitrary sequence that goes to infinity as* $n_{\min}\to\infty$*.*

Next by using Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (e.g. Massart, 1990), we extend the previous result to the randomized permutation test.

Corollary 5.5.1 (Uniform consistency of the randomized permutation test)**.** *Assume that* **(B1)** *and* **(B2)** *are fulfilled. Denote the Monte-Carlo-based permutation test function by* $\phi_{K,\mathrm{MC}}=\mathbb{1}(p_{\mathrm{MC}}\leq\alpha)$ *where* p_{MC} *is given in (5.6). Then under* H_1*,*

$$\lim_{M\to\infty}\limsup_{n_{\min}\to\infty}\sup_{(P_1,\ldots,P_K)\in\mathcal{F}_h(b_N r_N^\star)}\mathbb{P}\left(\phi_{K,\mathrm{MC}}=0\right)=0,$$

where b_N *is an arbitrary sequence that goes to infinity as* $n_{\min}\to\infty$*.*

Remark 5.8. *It is worth pointing out that the results of both Theorem 5.5 and Corollary 5.5.1 hold regardless of whether* K *is fixed or increases with* n_{min}*. However, we note that* K *cannot increase much faster than* $e^{n_{min}}$ *as* $\max_{1\leq k<l\leq K}\mathcal{V}_{kl}$ *is upper bounded by a positive constant under* ***(B1)*** *and thereby* $r_N^\star=\sqrt{\log K/n_{min}}$ *is also bounded.*

5.6.2 Minimax rate optimality

Theorem 5.5 as well as Corollary 5.5.1 show that the original and randomized permutation tests can be uniformly powerful over $\mathcal{F}_h(b_N r_N^\star)$ when b_N is sufficiently large. In this subsection, we focus on the MMD associated with a translation invariant kernel defined on $\mathbb{R}^d$ and further show that the previous result cannot be improved from a minimax point of view. A kernel $h:\mathbb{R}^d\times\mathbb{R}^d\mapsto\mathbb{R}$ is called *translation invariant* if there exists a symmetric positive definite function $\varphi:\mathbb{R}^d\mapsto\mathbb{R}$ such that $\varphi(x-y)=h(x,y)$ for all $x,y\in\mathbb{R}^d$ (Tolstikhin et al., 2017). Then our result is stated as follows.

Theorem 5.6. *Let $0 < \alpha < 1$ and $0 < \zeta < 1 - \alpha$. Suppose that $n_{\min} \to \infty$ and $K \to \infty$. Consider the class of sparse alternatives $\mathcal{F}_h(\epsilon_N)$ defined with a translation invariant kernel h on $\mathbb{R}^d$. Assume that there exists $z \in \mathbb{R}^d$ and $\kappa_1, \kappa_2 > 0$ such that $\varphi(0) - \varphi(z) \geq \kappa_1$ and $r_N^\star \leq \kappa_2$ for all $n_{\min}$. Further assume that* **(B1)** *and* **(B2)** *hold. Then under H_1, there exists a small constant $b > 0$ such that*

$$\liminf_{n_{\min} \to \infty} \inf_{\phi \in \Phi_N(\alpha)} \sup_{(P_1, \ldots, P_K) \in \mathcal{F}_h(br_N^\star)} \mathbb{P}(\phi = 0) \geq \zeta,$$

where $\Phi_N(\alpha)$ is the set of all level α test functions such that $\phi : \{Z_1, \ldots, Z_N\} \mapsto \{0, 1\}$.

Remark 5.9. *The results in Theorem 5.5 and Theorem 5.6 imply that the proposed permutation test is not only consistent but also minimax rate optimal against the considered sparse alternatives. As far as we are aware, this is the first time that the power of the permutation test is theoretically analyzed under large N and large K situations.*

Remark 5.10. *In our problem setup, a distance between two distributions is measured in terms of the maximum mean discrepancy associated with kernel h. One can also study minimax optimality of the proposed test over a class of alternatives measured in terms of a more standard metric such as the L_2 distance. For this direction, the results of Li and Yuan (2019) seem useful in which the authors explore minimax rate optimality of kernel mean embedding methods over a Sobolev space in the L_2 distance. We leave a detailed analysis of minimax optimality of the proposed test in other metrics to future work.*

5.7 Simulations

In this section, we demonstrate the finite-sample performance of the proposed approach via simulations. We consider two characteristic kernels for our test statistic; 1) Gaussian kernel and 2) energy distance kernel. Gaussian kernel is given by $h(x, y) = \exp(-\|x - y\|_2^2/\sigma)$ for which we choose the tuning parameter σ by the median heuristic (Gretton et al., 2012). On the other hand, energy distance kernel is given by $h(x, y) = (\|x\|_2 + \|y\|_2 - \|x - y\|_2)/2$ as before. Note that the MMD statistic with energy distance kernel is equivalent to the energy statistic (Székely and Rizzo, 2004; Baringhaus and Franz, 2004) in the two-sample case.

5.7.1 Other multivariate K-sample tests

We compare the performance of the proposed tests with two multivariate K-sample tests. The first one is the test based on DISCO statistic proposed by Rizzo and Székely (2010). Let $E_{kl,\alpha'}$ be the α'-energy statistic

between P_k and P_l given by

$$E_{kl,\alpha'} = \frac{2}{n_k n_l}\sum_{i_1=1}^{n_k}\sum_{i_2=1}^{n_l} g_{\alpha'}(X_{i_1,k}, X_{i_2,l}) - \frac{1}{n_k^2}\sum_{i_1,i_2=1}^{n_k} g_{\alpha'}(X_{i_1,k}, X_{i_2,k}) - \frac{1}{n_l^2}\sum_{i_1,i_2=1}^{n_l} g_{\alpha'}(X_{i_1,l}, X_{i_2,l}),$$

where $g_{\alpha'}(x,y) = \|x-y\|_2^{\alpha'}$. Let us write the between-sample and within-sample dispersions by $S_{\alpha'} = K^{-1}\sum_{1\le k<l\le K} E_{kl,\alpha'}$ and $W_{\alpha'} = 2^{-1}\sum_{k=1}^{K} n_k^{-1}\sum_{i_1,i_2=1}^{n_k} g_{\alpha'}(X_{i_1,k}, X_{i_2,k})$. Then DISCO statistic is defined as ratio of the between-sample dispersion to the within-sample dispersion, that is

$$D_\gamma = \frac{S_{\alpha'}/(K-1)}{W_{\alpha'}/(N-K)}.$$

The second test, proposed by Hušková and Meintanis (2008), is based on the empirical characteristic functions. For a given $\alpha'' \in \mathbb{R}$, Hušková and Meintanis (2008) consider the weighted L_2 distance between empirical characteristic functions as their test statistic, that is

$$H_{\alpha''} = \sum_{k=1}^{K}\frac{N-n_k}{Nn_k}\sum_{i_1,i_2=1}^{n_k} e^{-\|X_{i_1,k}-X_{i_2,k}\|_2^2/4\alpha''} - \frac{1}{N}\sum_{1\le k\ne l\le K}\sum_{i_1=1}^{n_k}\sum_{i_2=1}^{n_l} e^{-\|X_{i_1,k}-X_{i_2,l}\|_2^2/4\alpha''}.$$

In their paper, Hušková and Meintanis (2008) consider $\alpha'' = 1, 1.5, 2$ in their simulation study. Throughout our simulations, we choose $\alpha' = 1$ for $D_{\alpha'}$ and $\alpha'' = 1.5$ for $H_{\alpha''}$ and reject the null for large values of $D_{\alpha'}$ and $H_{\alpha''}$.

We also attempted to consider the graph-based K-sample test recently developed by Mukhopadhyay and Wang (2018). To implement their test, we used the R package provided by the same authors. Unfortunately, their method was not applicable when K is large due to numerical overflow in computing orthogonal polynomials. Hence we focus on the first two methods described in this subsection and compare them with the proposed tests against sparse alternatives.

5.7.2 Set-up

Let us denote a multivariate normal distribution with mean vector μ and covariance matrix Σ by $N(\mu,\Sigma)$. Similarly we denote a multivariate Laplace distribution with mean vector μ and covariance matrix Σ by $L(\mu,\Sigma)$. We examine the performance of the considered tests under the following sparse alternatives:

(a) **Normal Location**: $P_1 = N(\delta_1, I_d)$ and $P_2 = \ldots = P_K = N(\delta_0, I_d)$,

(b) **Normal Scale**: $P_1 = N(\delta_0, 3\times I_d)$ and $P_2 = \ldots = P_K = N(\delta_0, I_d)$,

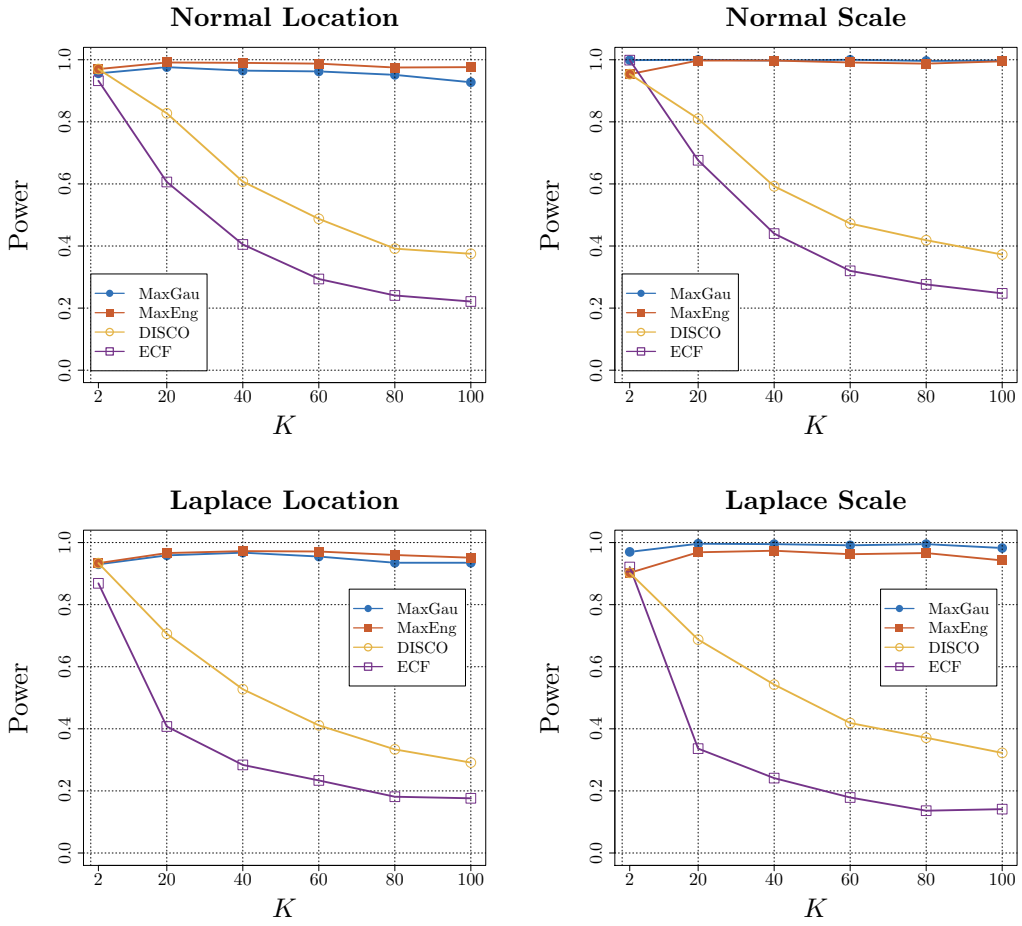

Figure 5.2: Empirical power comparisons of the considered tests against (a) Normal location, (b) Normal scale, (c) Laplace location, (d) Laplace scale alternatives. We refer to the tests based on $\widehat{\mathcal{V}}_{h,\max}$ with Gaussian kernel and energy distance kernel as MaxGau and MaxEng, respectively. In addition, the tests based on $D_{\alpha'}$ and $H_{\alpha''}$ are referred to as DISCO and ECF, respectively. See Section 8.9 for details.

(c) **Laplace Location**: $P_1 = L(\delta_{1.2}, I_d)$ and $P_2 = \ldots = P_K = L(\delta_0, I_d)$,

(d) **Laplace Scale**: $P_1 = L(\delta_0, 3 \times I_d)$ and $P_2 = \ldots = P_K = L(\delta_0, I_d)$,

where $\delta_b = (b, \ldots, b)^\top$ and I_d is the d-dimensional identity matrix. In words, we consider the sparse alternatives where only one of the distributions differs from the other $K-1$ distributions. Consequently, the signal is getting sparser as K increases. Throughout our experiments, we fix sample sizes $n_1 = n_2 = \ldots = n_K = 10$ and dimension $d = 5$ while increasing the number of distributions $K \in \{2, 20, 40, 60, 80, 100\}$. All tests were implemented via the randomized permutation procedure with $M = 200$ random permutations using the p-value in (5.6). As a result, they are all valid level α tests. Simulations were repeated 800 times to estimate the power at significance level $\alpha = 0.05$.

5.7.3 Results

From the results presented in Figure 5.2, we observe that the tests based on $D_{\alpha'}$ and $H_{\alpha''}$ have consistently decreasing power as K increases in all sparse scenarios. This can be explained by the fact that $D_{\alpha'}$ and $H_{\alpha''}$ are defined as an average between pairwise distances. Under the given sparse scenario, the average of pairwise distances, which is a signal to reject H_0, decreases as K increases. Hence the resulting tests based on $D_{\alpha'}$ and $H_{\alpha''}$ suffer from low power in large K. On the other hand, the proposed tests show robust performance to the number of distributions K under the given setting. They in fact have power very close to one even when K is considerably large, which emphasizes the benefit of using the maximum-type statistic against sparse alternatives.

Despite their good performance over sparse alternatives, the proposed tests do not always perform better than the average-type tests based on $D_{\alpha'}$ and $H_{\alpha''}$. For example, these average-type tests may outperform the proposed maximum-type tests against dense alternatives where many of $P_1, \ldots, P_K$ differ from each other. Given that prior knowledge on alternatives is not always available to users, developing a powerful test against both dense and sparse alternatives is an interesting direction for future work.

5.8 Conclusions

In this chapter, we introduced a new nonparametric K-sample test based on the maximum mean discrepancy. The limiting distribution of the proposed test statistic was derived based on Cramér-type moderate deviation for degenerate two-sample V-statistics. Unfortunately, the limiting distribution relies on an infinite number of nuisance parameters, which are intractable in general. Due to this challenge, we considered the permutation approach to determine the cut-off value of the test. We provided a concentration inequality for the proposed test statistic with a sharp exponential tail bound under permutations. On the basis of this result, we studied the power of the permutation test in large K and large N situations and further proved its minimax rate optimality under some regularity conditions. From our simulation studies, the proposed test is shown to be powerful against sparse alternatives where the previous methods suffer from low power. These findings suggest that our method will be useful in application areas where only a small number of populations differ from the others.

The power analysis in Section 5.6 relies on the assumption that a kernel is uniformly bounded. Although some of the popular kernels satisfy this assumption, our result cannot be applied to unbounded cases. One possible way to address this issue is to impose appropriate moment conditions on a kernel and utilize a suitable concentration inequality (e.g. a modified McDiarmid's inequality in Kontorovich, 2014) to obtain a similar result to Lemma 5.4.1. This topic is reserved for future work.

Chapter 6

Euclidean and Manhattan Distance for High-Dimensional Two-Sample Testing

This chapter is adapted from my work supervised by Sivaraman Balakrishnan and Larry Wasserman. While finishing up the work in August 2018, we found a similar paper appeared on ArXiv (June 2018) and recently published in *Stat* (Sarkar and Ghosh, 2018). Since significant parts of these two papers are overlapped and admittedly Sarkar and Ghosh (2018) present more solid results, we decided not to make this work available online.

6.1 Introduction

In recent years, high-dimensional data has become increasingly frequent in diverse fields of sciences. The rise of high-dimensional data has posed new challenges to traditional statistical methods. One challenge arises from the phenomenon of distance concentration in which all pairwise distances are almost equal in a high-dimensional space. In such cases, the notion of closeness may not be meaningful, which results in poor performance of distance based statistical methods. This phenomenon was examined by many authors such as Aggarwal et al. (2001), Hall et al. (2005), Ahn et al. (2007), Sarkar and Ghosh (2019) in the context of classification, clustering and dimension reduction.

The concept of closeness is also crucial in two-sample testing problems. A number of testing procedures rely on a certain notion of proximity and Euclidean distance is among the widely used metric in the literature. The existing distance-based two-sample procedures can be summarized into two categories. One approach

is based on a comparison between within-class distances and between-class distances and the null is rejected when the distributions of these distances do not coincide. The testing procedures introduced by Baringhaus and Franz (2004), Székely and Rizzo (2004), Bakshaev (2009) and Biswas and Ghosh (2014) can fall into this category. Another approach is based on a graph construction. In this approach, we construct a graph based on sample distances and count how many edges are connected according to a given rule of the test. The graph-based approach includes the tests based on the k-nearest neighbor (Schilling, 1986; Henze, 1988), the minimum spanning tree (Friedman and Rafsky, 1979), the non-bipartite matching (Rosenbaum, 2005) and the shortest Hamiltonian path (Biswas et al., 2014).

When Euclidean distance is used as a measure of proximity, all of the aforementioned tests have been shown to be consistent against general alternatives in low-dimensional settings. In other words, the asymptotic power of these tests converge to one for any difference between two distributions. In high-dimensions, however, this general consistency is no longer obvious due to the phenomenon of distance concentration.

In this work, we focus on the regime where the dimension tends to infinity while the sample size is fixed. In this high-dimension, low sample size setting, we show that the Euclidean-based two-sample tests can be consistent only against first or second moment differences under weakly dependent conditions on random vectors. To overcome this problem, we suggest Manhattan distance as an alternative to Euclidean distance. We then show that the nonparametric tests based on Manhattan distance are consistent against a broader range of alternatives than those based on Euclidean distance under the high-dimensional regime.

6.2 Motivating Example

Suppose that $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ are d-dimensional random vectors independently from two distributions F_X and G_Y, respectively. We write $X_i = (X_{i1}, \ldots, X_{id})^\top$ and $Y_j = (Y_{j1}, \ldots, Y_{jd})^\top$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Based on these samples, the two-sample problem is concerned with testing whether $H_0 : F_X = G_Y$ or $H_1 : F_X \neq G_Y$. For an illustrative example, assume that the components of X_1 are independent and identically distributed as $N(0, 1)$. Similarly assume that the components of Y_1 are independent and identically distributed as $\lambda N(\mu_1, \sigma_1^2) + (1 - \lambda)N(\mu_2, \sigma_2^2)$. We then set $\lambda = 0.5$, $\sigma_1^2 = \sigma_2^2 = 0.2$, $\mu_1 = (1 - \sigma_1^2)^{1/2}$, and $\mu_2 = -\mu_1$. In this setting, the underlying two distributions have the same mean and the same covariance matrix, but they clearly have different higher moments. To illustrate the problem of high-dimensional Euclidean distance, we consider the three nonparametric tests commonly used in the literature. The first one is the interpoint distance-based test proposed by Baringhaus and Franz (2004), and the second one is based on the k-nearest neighbor graph by Schilling (1986) where we chose $k = 3$ in our simulation study. The third one is a multivariate generalization of the run test by Friedman and Rafsky (1979). The definition of each test is provided in Section 6.3. For the simulation study, we

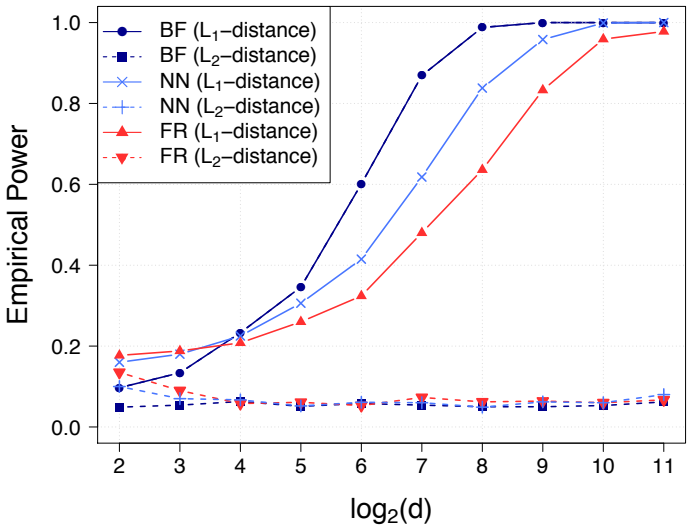

Figure 6.1: Power comparison between the Baringhaus and Franz (BF) test, the nearest neighbor (NN) test and the Friedman and Rafsky (FR) test at significant level $\alpha = 0.05$. For each test, we considered Euclidean distance and Manhattan distance to illustrate their different behaviors in the high-dimensional regime.

let the sample size $m = n = 20$, and increased the dimension to describe high-dimensional behavior of the considered tests. We used the permutation procedure with 500 permutations to decide a cut-off value of each test, and the simulations were repeated 1,000 times to estimate the power.

The simulation result is provided in Figure 6.1. Under the considered scenario, it is natural to expect that the power of any reasonable nonparametric test increases with the dimension since we gain more evidence against the null hypothesis. However, this was not the case for the tests based on Euclidean distance. They perform poorly over the range of dimensions. On the other hand, the power of the tests based on Manhattan distance consistently increase to one with the dimension. Understanding this counterintuitive result is the main object of this work.

6.3 The Problem of High-Dimensional Euclidean Distance

The high-dimensional behavior of Euclidean distance has been investigated by several authors (e.g., Hall et al., 2005; Biswas et al., 2014; Mondal et al., 2015). To summarize the previous analysis, assume that there exist $\sigma_x, \sigma_y > 0$ such that $d^{-1}\sum_{i=1}^{d} \mathrm{var}(X_{1i}) \to \sigma_x^2$ and $d^{-1}\sum_{i=1}^{d} \mathrm{var}(Y_{1i}) \to \sigma_y^2$. Further assume $d^{-1}\sum_{i=1}^{d}\{E(X_{1i}) - E(Y_{1i})\}^2 \to \tau_{xy}^2$ as $d \to \infty$. Then under the appropriate assumptions on the moments and the dependent structure of X and Y (see e.g., Biswas et al., 2014), the weak law of large numbers shows that $d^{-1/2}||X_1 - X_2||_2 = d^{-1/2}\{\sum_{i=1}^{d}(X_{1i} - X_{2i})^2\}^{1/2}$ converges to $\sigma_x\sqrt{2}$ in probability. Similarly,

$d^{-1/2}||Y_1 - Y_2||_2$ and $d^{-1/2}||X_1 - Y_1||_2$ converge in probability to $\sigma_y\sqrt{2}$ and $(\sigma_x^2 + \sigma_y^2 + \tau_{xy}^2)^{1/2}$, respectively. This implies that any pairwise Euclidean distance between samples from the same class is approximately identical as either $\sigma_x\sqrt{(2d)}$ or $\sigma_y\sqrt{(2d)}$, depending on a given class. Similarly, any pairwise Euclidean distance between samples from the different classes is of nearly equal length as $\{d(\sigma_x^2 + \sigma_y^2 + \tau_{xy}^2)\}^{1/2}$. Intuitively, in order for Euclidean-based tests to achieve reasonable power in the high-dimensional setting, at least one of the quantities among $\sigma_x\sqrt{2}$, $\sigma_y\sqrt{2}$, and $(\sigma_x^2 + \sigma_y^2 + \tau_{xy}^2)^{1/2}$ should be distinct from the others. However, in the cases where two distributions have the same mean and the same covariance matrix, these quantities are all identical, and the resulting Euclidean-based tests have low power as in the previous motivating example.

Although the above argument based on the convergence in probability is intuitive, it only tells us that the sample Euclidean distance is concentrated around its population quantity as the dimension tends to infinity. To have a deeper understanding of the given argument, we investigate the limiting distribution of Euclidean distance in the high-dimensional setting. To begin with, we introduce a α-mixing condition.

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two σ-fields. The α-mixing coefficient between $\mathcal{G}_1$ and $\mathcal{G}_2$ is defined by

$$\alpha(\mathcal{G}_1, \mathcal{G}_2) = \sup_{A \in \mathcal{G}_1, B \in \mathcal{G}_2} |\text{pr}(A \cap B) - \text{pr}(A)\text{pr}(B)|.$$

We denote the σ-field generated by the sequence of random variables $\{X_{1i}\}_{i=l}^k$ by $\mathcal{F}_l^k$. Then the α-mixing coefficient of $\{X_{1i}\}_{i=1}^\infty$ is given by

$$\alpha_X(r) = \sup_{k \geq 1} \alpha(\mathcal{F}_1^k, \mathcal{F}_{r+k}^\infty).$$

The sequence $\{X_{1i}\}_{i=1}^\infty$ is called α-mixing if $\lim_{r\to\infty} \alpha_X(r) = 0$. We refer to Lin and Lu (1997) and Chapter 16.2 of Athreya and Lahiri (2006) for more details about α-mixing sequences.

Let $\{Z_i\}_{i=1}^N$ be the combined samples of $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{i=1}^n$ where $N = m + n$. We denote all possible pairwise Euclidean distances between $\{Z_i\}_{i=1}^N$ by

$$(W_1, \ldots, W_M)^\top = (||Z_1 - Z_2||_2, \ldots, ||Z_{N-1} - Z_N||_2)^\top,$$

where $M = N(N-1)/2$. With these notations, we establish the multivariate central limit theorem for the pairwise Euclidean distances.

Lemma 6.0.1. *Let $\{X_{1i}\}_{i=1}^\infty$ and $\{Y_{1i}\}_{i=1}^\infty$ be strictly stationary sequences with $E(X_{11}) = \mu_x, \text{var}(X_{11}) = \sigma_x^2$ and $E(Y_{11}) = \mu_y, \text{var}(Y_{11}) = \sigma_y^2$. For some $\delta \in (0, \infty)$, suppose that following assumptions hold:*

1. *The moments $E|X_{11}|^{4+2\delta}$ and $E|Y_{11}|^{4+2\delta}$ are bounded.*

2. *The α-mixing coefficients of $\{X_{1i}\}_{i=1}^{\infty}$ and $\{Y_{1i}\}_{i=1}^{\infty}$ satisfy*

$$\sum_{r=1}^{\infty} \alpha_X(r)^{\delta/2+\delta} < \infty, \quad \sum_{r=1}^{\infty} \alpha_Y(r)^{\delta/2+\delta} < \infty.$$

3. *The minimum eigenvalue of $\lim_{d\to\infty} \mathrm{var}\{d^{-1/2}(W_1^2, \ldots, W_M^2)^\top\}$ is positive.*

Then

$$(W_1, \ldots, W_M)^\top - (\mu_1, \ldots, \mu_M)^\top$$

converges to the multivariate normal distribution with zero mean vector and positive definite covariance matrix Σ. Here μ_i is one of the values among $\sigma_x\sqrt{(2d)}$, $\sigma_y\sqrt{(2d)}$ and $\{d\sigma_x^2 + d\sigma_y^2 + d(\mu_x - \mu_y)^2\}^{1/2}$.

Remark 6.1. *The strictly stationary condition in Lemma 6.0.1 can be removed by using the multivariate central limit theorem for non-stationary dependent random sequences with extra conditions (e.g. Theorem 16.3.5 of Athreya and Lahiri, 2006).*

Using Lemma 6.0.1, we demonstrate that the nonparametric tests based on Euclidean distance do not have consistency against general alternatives in the high-dimensional regime. Specifically, we show under the assumptions in Lemma 6.0.1 that the power of the tests do not converge to one unless there exist first or second moment differences. We simply focus on Friedman and Rafsky's test (Friedman and Rafsky, 1979), the nearest neighbor test (Schilling, 1986) and Baringhaus and Franz's test (Baringhaus and Franz, 2004) considered in the previous motivating example, but the same argument can be similarly applied to other nonparametric tests based on Euclidean distance. Friedman and Rafsky's test rejects the null hypothesis for a small value of its statistic, which is defined by

$$T_{m,n}^{\mathrm{FR}} = \sum_{i=1}^{N-1} \Psi_i + 1.$$

Here Ψ_i is the indicator variable equal to one if and only if the ith edge of the minimal spanning tree connects two observations from the different distributions. On the other hand, the nearest neighbor test rejects the null hypothesis when its statistic is larger than a cut-off value. The nearest neighbor statistic is defined by

$$T_{m,n}^{\mathrm{NN}} = \frac{1}{kN} \sum_{i=1}^{N} \sum_{r=1}^{k} I_i(r),$$

where $I_i(r)$ is the indicator variable equal to one if and only if X_i and its rth nearest neighbor are from the same distribution. Lastly, Baringhaus and Franz's test rejects the null hypothesis for a large value of its

statistic and its test statistic is given by

$$T_{m,n}^{\mathrm{BF}} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}||X_i - Y_j||_2 - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}||X_i - X_j||_2 - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}||Y_i - Y_j||_2.$$

Based on the given tests, we have the following result.

Theorem 6.1. *Suppose F_X and G_Y have the same mean and the same covariance matrix, but different higher moments. Under the assumptions in Lemma 6.0.1, the power of $T_{m,n}^{FR}$, $T_{m,n}^{NN}$ and $T_{m,n}^{BF}$ based on Euclidean distance do not converge to one for any choice of significance level $\alpha \in (0,1)$ as $d \to \infty$.*

Under stronger assumptions, we can show that certain types of Euclidean-based tests become completely powerless under the high-dimensional settings. The result is stated in the following theorem.

Theorem 6.2. *Let the components of X_1 be independent and identically distributed. Similarly, let the components of Y_1 be independent and identically distributed as well. Suppose that F_X and G_Y have the same finite moments up to their fourth moments. In other words, we have $E|X_{11}|^p = E|Y_{11}|^p < \infty$ for $p = 1, 2, 3, 4$. Consider a function $g : \mathbb{R}^M \mapsto \mathbb{R}$ that has a derivative at $(2\sigma_x^2, \ldots, 2\sigma_x^2)$ where $\sigma_x^2 = \mathrm{var}(X_{11})$ and define a test statistic as*

$$T_{m,n} = g\left(W_1, \ldots, W_M\right). \tag{6.1}$$

Suppose that we reject the null if $T_{m,n} \in R_\alpha$ where R_α is a α level rejection region. Then the power of the considered test becomes less than or equal to α as $d \to \infty$.

The examples of the test statistic that has the form of (6.1) include the test statistics proposed by Baringhaus and Franz (2004), Székely and Rizzo (2004), Biswas and Ghosh (2014) and the kernel maximum mean discrepancy with the radial basis kernel by Gretton et al. (2012).

6.4 Alternative approach based on Manhattan Distance

To address the problem of Euclidean-based two-sample tests, we consider Manhattan distance as an alternative to Euclidean distance. In particular, we illustrate that the tests based on Manhattan distance are consistent against more general alternatives than those based on Euclidean distance in the high-dimensional setting. We begin by introducing some notations.

Let $F_{X_{11}}(t)$ and $G_{Y_{11}}(t)$ be the distributions of X_{11} and Y_{11}, respectively. We write

$$\gamma_x = 2\int_{-\infty}^{\infty} F_{X_{11}}(t)\left(1 - F_{X_{11}}(t)\right)dt,$$

$$\gamma_y = 2\int_{-\infty}^{\infty} G_{Y_{11}}(t)\,(1 - G_{Y_{11}}(t))\,dt,$$
$$\gamma_{xy} = \int_{-\infty}^{\infty} G_{Y_{11}}(t)\,(1 - F_{X_{11}}(t))\,dt + \int_{-\infty}^{\infty} F_{X_{11}}(t)\,(1 - G_{Y_{11}}(t))\,dt.$$

Then we provide the following lemma.

Lemma 6.2.1. *Consider the same assumptions in Lemma 6.0.1. Then the Manhattan distance between X_1 and X_2 scaled by d, that is $d^{-1}||X_1 - X_2||_1 = d^{-1}\sum_{i=1}^{d}|X_{1i} - X_{2i}|$, converges to γ_x in probability. Similarly, $d^{-1}||Y_1 - Y_2||_1$ and $d^{-1}||X_1 - Y_1||_1$ converge in probability to γ_y and γ_{xy}, respectively. In addition, we have $\gamma_x = \gamma_y = \gamma_{xy}$ if and only if the marginal distributions of X and Y are the same.*

The above lemma shows that if two distributions have different marginal distributions, at least one of γ_x, γ_y and γ_{xy} is distinct from the others. This result differs from that of Euclidean distance whose limit only depends on the first two moments. As a result, Manhattan-based tests can be sensitive against higher moment alternatives where Euclidean-based tests become powerless. In the following, we revisit the normal mixture example in Section 6.2 and further illustrate our main point.

Example 6.1 (Normal mixture)**.** *Let us write the mean of a folded normal random variable with parameters (μ, σ^2) as a function of μ and σ^2 by*

$$f(\mu, \sigma^2) = \sigma\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left\{1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right\},$$

where Φ is the standard normal cumulative distribution function. Then under the normal mixture example given in Section 6.2, the limits of pairwise Manhattan distances γ_x, γ_y and γ_{xy} are calculated as follows (the details are presented in the supplementary material):

$$\gamma_x = \frac{2}{\sqrt{\pi}}, \quad \gamma_y = \lambda^2 f\left(0, 2\sigma_1^2\right) + 2\lambda(1-\lambda) f\left(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2\right) + (1-\lambda)^2 f\left(0, 2\sigma_2^2\right),$$
$$\gamma_{xy} = \lambda f(\mu_1, 1 + \sigma_1^2) + (1-\lambda) f(\mu_2, 1 + \sigma_2^2).$$

By plugging in $\lambda = 0.5, \sigma_1^2 = \sigma_2^2 = 0.2, \mu_1 = \sqrt{1 - \sigma_1^2}$ and $\mu_2 = -\mu_1$, the limit values are approximated by $\gamma_x \approx 1.128$, $\gamma_y \approx 1.147$ and $\gamma_{xy} \approx 1.150$. Therefore, γ_x, γ_y and γ_{xy} are all distinct and $\gamma_{xy} > \gamma_y > \gamma_x$. This result contrasts with the limit values of pairwise Euclidean distances, which are all identical under the given scenario, and therefore explains the superior behavior of the Manhattan-based tests in Figure 6.1.

In the next theorem, we briefly describe the cases where three tests considered in Theorem 6.1 are consistent when Euclidean distance is replaced with Manhattan distance. However, they can be consistent against other cases supported by our simulation results.

Theorem 6.3. *Assume the same scenario in Theorem 6.1 where the Euclidean-based tests fail to be consistent. Further assume that* $\gamma_{xy} > \max\{\gamma_x, \gamma_y\}$. *In this case, the Manhattan-based tests have the following asymptotic behavior in the high-dimensional setting:*

(i). For any $\alpha > \max\{\lfloor N/n \rfloor, \lfloor N/m \rfloor\}/\{N!/(m!n!)\}$, *the power of Friedman and Rafsky's test converges to one as* $d \to \infty$.

(ii). For any $\alpha > [1 + n!/\{m!(m-n)!\}]/\{N!/(m!n!)\}$ *and* $m/2 \leq k < m$ *where* $2 \leq m \leq n$, *the power of the nearest neighbor test converges to one as* $d \to \infty$.

(iii). For any $\alpha > 1/\{N!/(m!n!)\}$ *when* $m \neq n$ *and* $\alpha > 2/\{N!/(m!n!)\}$ *when* $m = n$, *the power of Baringhaus and Franz's test converges to one as* $d \to \infty$.

Remark 6.2. *Suppose that there exists* $\varepsilon \in (0, 1]$ *such that* $m = \varepsilon n$. *Then the lower bounds for significance level* α *in Theorem 6.3 can be arbitrary small by choosing* n *sufficiently large.*

6.5 Simulations

In this section, we provide simulation results for the tests based on Euclidean and Manhattan distance against several alternatives. We consider Friedman and Rafsky's test, the nearest neighbor test, Baringhaus and Franz's test and their variants for comparison. Recent studies (Mondal et al., 2015; Chen and Friedman, 2017) show that the nearest neighbor test and Friedman and Rafsky's test perform poorly against scale differences in high-dimensions. Hence, in our simulation study, we also consider the modified tests proposed by Mondal et al. (2015) and Chen and Friedman (2017). First, the modified nearest neighbor test statistic (Mondal et al., 2015) is

$$T_{m,n}^{\mathrm{MBG}} = \frac{m}{N}\left(T_{m,k} - \frac{m-1}{N-1}\right)^2 + \frac{n}{N}\left(T_{n,k} - \frac{n-1}{N-1}\right)^2, \tag{6.2}$$

where $T_{m,k} = \sum_{i=1}^{m}\sum_{r=1}^{k} I_i(r)/(mk)$ and $T_{n,k} = \sum_{i=1}^{n}\sum_{r=1}^{k} I_i(r)/(nk)$. As before, we chose $k = 3$ as the number of neighbors for the simulation study. On the other hand, Chen and Friedman (2017) proposed modifications of Friedman and Rafsky's test. Let R_m be the number of edges connecting two observations from F_X and let R_n be the number of edges connecting two observations from G_Y. Among the different tests that they proposed, we focus on the test based on the following statistic:

$$T_{m,n}^{\mathrm{CF}} = \left\{R_m - |G|\frac{m(m-1)}{N(N-1)}\right\}^2 + \left\{R_n - |G|\frac{n(n-1)}{N(N-1)}\right\}^2, \tag{6.3}$$

where $|G|$ is the number of edges in the minimum spanning tree.

Table 6.1: Empirical power of the tests over different dimensions at significance level $\alpha = 0.05$ and $m = n = 20$.

		Location			*Scale*			*Kurtosis*		
	d	100	500	1000	100	500	1000	100	500	1000
$T^{\mathrm{BF}}_{m,n}$	*Manhattan*	0.701	1.000	1.000	0.084	0.178	0.218	**0.752**	**1.000**	**1.000**
	Euclidean	0.748	1.000	1.000	0.096	0.178	0.239	0.058	0.050	0.050
$T^{\mathrm{NN}}_{m,n}$	*Manhattan*	0.412	0.897	0.988	0.068	0.004	0.000	**0.543**	**0.969**	**1.000**
	Euclidean	0.453	0.936	0.998	0.065	0.002	0.000	0.058	0.054	0.056
$T^{\mathrm{FR}}_{m,n}$	*Manhattan*	0.312	0.733	0.928	0.059	0.000	0.000	**0.433**	**0.843**	**0.960**
	Euclidean	0.342	0.752	0.946	0.047	0.000	0.000	0.069	0.070	0.050
$T^{\mathrm{MBG}}_{m,n}$	*Manhattan*	0.069	0.271	0.591	0.871	1.000	1.000	**0.157**	**0.676**	**0.967**
	Euclidean	0.082	0.270	0.635	0.871	1.000	1.000	0.090	0.104	0.086
$T^{\mathrm{CF}}_{m,n}$	*Manhattan*	0.103	0.276	0.461	0.749	1.000	1.000	**0.202**	**0.586**	**0.853**
	Euclidean	0.094	0.250	0.500	0.755	0.999	1.000	0.117	0.095	0.115

We compare the performance of the tests against the three alternatives where differences are in location, scale and kurtosis parameters, respectively. For the location alternative, we let F_X and G_Y be the multivariate normal distributions as $N_d((0.2,\ldots,0.2)^\top, I)$ and $N_d((0,\ldots,0)^\top, I)$. Similarly, for the scale alternative, we choose $F_X = N_d((0,\ldots,0)^\top, I)$ and $G_Y = N_d((0,\ldots,0)^\top, 1.1^2 \times I)$. For the kurtosis alternative, we reconsider the multivariate normal mixture example described in Section 6.2 where the two distribution have the same mean vector and the same covariance matrix. The simulation results under these scenarios are provided in Table 6.1. All of the tests were implemented based on the permutation procedure with 500 permutations, and the simulations were repeated 1,000 times to estimate the power.

For the main comparison between Euclidean and Manhattan distance, we see that the Manhattan-based tests are comparable to or slightly less powerful than the Euclidean-based tests against the location and scale alternatives. However, we would like to emphasize that the Euclidean-based tests perform poorly against the kurtosis alternative for the reasons described in Section 6.3. In contrast, we observe that the power of the Manhattan-based tests keep increasing with d against the kurtosis alternative, which confirms our previous analysis. For the comparison between different two-sample methods, Baringhaus and Franz's test outperforms the other tests against the location and kurtosis alternatives, however it is less powerful than the test by Mondal et al. (2015) and the test by Chen and Friedman (2017) against the scale alternative. In general, the nearest neighbor test and the Friedman and Rafsky' test are less powerful than the other tests under the considered examples.

Chapter 7

Classification accuracy as a proxy for two-sample testing

This chapter is adapted from my joint work with Aaditya Ramdas, Aarti Singh and Larry Wasserman. This work is accepted to *the Annals of Statistics* for publication.

7.1 Introduction

The recent popularity of machine learning has resulted in the extensive teaching and utilization of prediction methods in theoretical and applied communities. When faced with a hypothesis testing problem in practice, data scientists sometimes opt for a prediction-based test-statistic.

We study one example of this common practice in this chapter, concerning arguably the most classical testing and prediction problems — *two-sample testing* (are the two underlying distributions the same?) and *classification* (learning a classifier that separates the two distributions, implicitly assuming they are not the same). Practitioners familiar with machine learning but not the hypothesis testing literature often find it intuitive to perform testing in the following way: first learn a classifier, and then see if its accuracy is significantly different from chance and if it is, then conclude that the distributions are different.

The central question that this chapter seeks to answer is "*what are the pros and cons of the classifier-based approach to two-sample testing?*". As we shall detail in Section 7.2, the notion of *cost* or *price* that is appropriate for the Neyman-Pearson or Fisherian hypothesis testing paradigm, is the power achievable at a fixed false positive level α (in other words, the lowest possible type-2 error achievable at some prespecified target type-1 error). Indeed, we approach this question using the frequentist perspective of minimax theory. More formally, we can restate our question as "*when is the classifier-based test consistent, and how does its power compare to the minimax power?*".

7.1.1 Practical motivation

Before we delve into the details, it is worth mentioning that even though this chapter is a theoretical endeavor, the question was initially practically motivated. Many scientific questions are naturally posed as two-sample tests — examples abound in epidemiology and neuroscience. As a hypothetical example from the latter, say we are interested in determining whether a particular brain region responds differently under two situations (say listening to loud harsh sounds vs soft smooth sounds), or for a person with a medical condition (patient) and a person without the condition (control). Often, one collects and analyzes brain data for the same patient under the two contrasting stimuli (to study the effect of change in that stimulus), or for different normal and ill patients under the same stimulus (to study effect of a medical condition). Since the work of Golland and Fischl (2003) where the authors examined permutation tests for classification with application to neuroimaging analysis, it has been increasingly common in the field of neuroscience—see Zhu et al. (2008); Etzel et al. (2009); Pereira et al. (2009); Stelzer et al. (2013)—to assess whether there is a significant difference between the two sets of data collected by learning a classifier to differentiate between them (because, for instance, they may be more familiar with classification than two-sample testing). Neuroscientists call this style of brain decoding as pattern discrimination and a positive answer can be seen as preliminary evidence that the mental process of interest might occur within the portion of the brain being studied; see Olivetti et al. (2012) for a discussion of related issues. This classification approach to two-sample testing has been considered in other application areas including genetics (Yu et al., 2007), speech analysis (Chen et al., 2009), credit scoring (Xiao et al., 2014), churn prediction (Xiao et al., 2015) and video content analysis (Liu et al., 2018).

7.1.2 Overview of the main results

Our first contribution is to identify weak conditions on the classifier that suffice for both finite-sample or asymptotic type-1 error control, as well as for asymptotic consistency.

- **Asymptotic test (Proposition 7.3):** We identify mild conditions under which the sample-splitting error of a general classifier (7.27) is asymptotically normal as $n, d \to \infty$. We introduce an asymptotic test based on this Gaussian approximation and prove its asymptotic type-1 error control. We also prove that a sufficient condition for its consistency (for its power to asymptotically approach one) is that its true accuracy converges to $1/2 + \epsilon$ for any constant $\epsilon > 0$ as $n, d \to \infty$ at any relative rate.

- **Permutation test (Theorem 7.6):** In addition to the asymptotic approach, we consider two types of random permutation procedures that yield a valid level α test in finite-sample scenarios. Under the same conditions made before, we present the minimum number of random permutations that guarantees that the resulting permutation test is consistent.

For technical reasons, it is most convenient to present these results last, after suitable notation, lemmas and assumptions have been developed in earlier sections.

The above results leave two natural questions open: first, whether we can derive a rate of consistency in special cases, and second, whether testing can be consistent even when the classifier accuracy approaches chance (is not bounded away from half). We answer both affirmatively; our second contribution is to rigorously analyze the asymptotic power of tests using classification accuracy for Gaussian and elliptical distributions in a high-dimensional setting when the error of the Bayes optimal classifier approaches half. In this direction, we have three main results:

- **Power of the accuracy of LDA for Gaussian distributions with known Σ (Theorem 7.3):** The considered test statistic (7.14) is the centered and rescaled classification error of LDA estimated via sample splitting, when Σ is known. Under standard interpretable assumptions (Section 7.5.1), this test statistic converges to a standard normal in the high-dimensional setting (Theorem 7.2) under both null and local alternative. Using this fact, we describe its local asymptotic power in expression (7.20). Comparing the latter with the minimax power (7.8), we highlight that the performance of the accuracy test is comparable to but worse than the minimax optimal test, achieving an asymptotic relative efficiency (ARE) of $1/\sqrt{\pi} \approx 0.564$ for balanced sample sizes.

- **Extensions to unknown Σ using naive Bayes and other variants (Theorem 7.4):** We generalize the previous findings to other linear classifiers for unknown Σ, like naive Bayes. We again find that classifier-based tests are underpowered, achieving the same aforementioned ARE of $1/\sqrt{\pi}$ compared to corresponding variants of Hotelling's test such as Bai and Saranadasa (1996) and Srivastava and Du (2008).

- **Extensions to elliptical distributions (Theorem 7.5):** We extend Theorem 7.3 to the class of elliptical distributions with finite kurtosis, and prove that the asymptotic power expression remains unchanged from the Gaussian setting, up to an explicit constant factor, which is $\sqrt{2}$ times the marginal density evaluated at 0. Restricting our attention to multivariate t-distributions, we also find an interesting phenomenon that the classifier-based test becomes relatively more efficient when the underlying distributions have heavier tails.

As two side contributions, we formally study the fundamental minimax power of high-dimensional two-sample mean testing for Gaussians. In this direction, we have two main results.

- **Explicit and exact expression for asymptotic minimax power (Proposition 7.1):** By building on prior work (Luschgy, 1982), we provide an explicit expression for the asymptotic minimax power of high-dimensional two-sample mean testing that is valid for any positive definite covariance matrix and

unbalanced sample sizes when $d, n \to \infty$ at any relative rate.

- **Minimax optimality of Hotelling's T^2 test when $d = o(n)$ (Theorem 7.1):** It is well known that Hotelling's test is minimax optimal when d is fixed and $n \to \infty$. In the high-dimensional setting, when the dimension d and the sample size n both increase to infinity with $d/n \to c \in (0, 1)$, Bai and Saranadasa (1996) show that Hotelling's test may have low power. Since then, Hotelling's test has been largely undervalued in the setting where d increases with n. In contrast to the aforementioned negative result, we prove that Hotelling's test remains asymptotically minimax optimal when $d \to \infty$ as long as $d/n \to 0$.

7.1.3 Interpreting our results and practical takeaway messages

There may be two somewhat contradictory ways that our results may be interpreted:

1. Practitioners may (possibly unjustly) use our results to reassure themselves that their utilization of relatively more flexible prediction methods for testing, even in the high dimensional setting, may not hurt their power too much.

2. At the same time, our results may also serve as a warning that a constant factor loss of power might be possible, and for scientific disciplines in which data is not abundant, the scientist may be wary of using prediction methods for hypothesis testing problems.

Indeed, after our earlier arXiv manuscript appeared, a few different papers have cited our results to justify their practical choices in both of these above ways. To help weigh in on this possible conundrum and stop the apparently contradictory messaging, we take the liberty of using our intuition from this chapter and also other recent papers (e.g. Lopez-Paz and Oquab, 2016; Hediger et al., 2019; Gagnon-Bartsch and Shem-Tov, 2019) to instead propose complementary, non-contradictory takeaway messages:

1. If the data is relatively unstructured or not abundant, and if the alternative can be accurately specified in such a manner that is both practically meaningful and for which a provably powerful two-sample test statistic is available (or can be easily designed), then we recommend using such a well-tailored statistic.

2. Suppose the data is highly structured or abundant (say, images of two species of beetles), but the potential differences between the two distributions cannot be easily specified. In this case, constructing a refined test that has high power against an accurately prespecified alternative may be too hard, and thereby we recommend using a flexible two-sample test statistic like classification accuracy (say using a convolutional neural network classifier or random forests).

Of course, it seems very challenging to theoretically study these setups in their full generality to provide a thorough formal backing to such practical suggestions. However, we are hopeful that our work will spur others to extend our concrete results to new settings.

7.1.4 Related work

The idea of using binary classifiers for two-sample testing was conceptualized by Friedman (2004). However, Friedman's proposal was fundamentally different from the one proposed here: he suggested using training a classifier on all points, and using that classifier to assign a score to each point, and the scores in each class were compared using a univariate two-sample test like Mann-Whitney or Kolmogorov-Smirnov. In other words, Friedman's proposal was to use classifiers to reduce a multivariate two-sample test into a univariate one. A different classifier-based approach to the two-sample problem was proposed by Blanchard et al. (2010). Although their test statistic is built upon classification algorithms, it aims to estimate the a priori probability of a contamination model, instead of classification accuracy.

This chapter in contrast considers held-out accuracy as the test statistic. The held-out accuracy of any classifier in any dimension can be used as the test statistic, and type-1 error can always be controlled non-asymptotically at the desired level using permutations. Hence, the main question of genuine mathematical interest is what we can prove about the power of such a test. Instead of permutations, if we instead use a Gaussian approximation to the null distribution, then it is unclear whether it remains valid in the high-dimensional setting and again its power is unclear. To the best of our knowledge, our Feb'16 ArXiv manuscript was the first mathematical attempt to study the power of this approach. There has been a growing interest in this idea in both the statistics and the machine learning communities (Rosenblatt et al., 2016; Lopez-Paz and Oquab, 2016; Borji, 2019; Hediger et al., 2019; Gagnon-Bartsch and Shem-Tov, 2019), most of which build on our ArXiv preprint but further provide valuable practical insight into the problem using various classifiers under different scenarios. Nevertheless most of these other works couple informal arguments with numerical experiments, motivating us to fully formalize and generalize our earlier analysis.

In an orthogonal work, Scott and Nowak (2005) proposed a Neyman-Pearson classification framework within which one would like to minimize the probability of classification error for one class, subject to a bound on the probability of classification error for the other class. Their problem is a variant of classification in which the classifier is judged by a different error metric, but it is quite different from our goal of two-sample testing. Other connections between classification and two-sample testing have also been explored by Ben-David et al. (2007), Fukumizu et al. (2009) and Gretton et al. (2012), but none of them set out to solve our problem.

Another class of two-sample tests is based on geometric graphs; examples include the k-nearest neighbor (NN) graph (Schilling, 1986; Henze, 1988), the minimum spanning tree (Friedman and Rafsky, 1979) and the

cross-matching (Rosenbaum, 2005). Recently Bhattacharya (2018) presented general asymptotic properties of graph-based tests under the fixed dimensional setting. Comparing the performance of the k-NN graph test and the k-NN classifier test (based on its heldout classification accuracy, as studied in this chapter) may be interesting to explore in future work.

There is of course a very large body of work that just analyzes classifiers, or just analyzing two-sample tests (e.g. Hu and Bai, 2016; Arias-Castro et al., 2018, and the references therein), but without connecting the two. These will be cited when their results are used.

Paper Outline. The rest of this chapter is organized as follows. In Section 7.2, we formally define both testing and classification problems. In Section 7.3, we discuss a minimax lower bound for two-sample testing in high-dimensional settings and in Section 7.4, we prove that Hotelling's T^2 test achieves this lower bound when $d/n \to 0$. Section 7.5 studies the limiting distribution of Fisher's LDA accuracy in the high-dimensional setting. Building on this limiting distribution, Section 7.6 presents the asymptotic power of Fisher's LDA for two-sample mean testing under known Σ. Section 7.7 extends this asymptotic power expression to other linear classifiers with unknown Σ, like naive Bayes. Generalizations to elliptical distributions are in Section 7.8. In Section 7.9, we examine the type-1 error control and consistency of the asymptotic test as well as the permutation test for *any* classifier. In Section 7.10, we provide simulation results that confirm our theoretical analysis, before concluding in Section 7.11. The proofs of all the results along with the discussion on open problems are provided in the supplement.

Notation. Let $\mathcal{N}_d(\mu, \Sigma)$ refer to the d-variate Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and $d \times d$ positive definite covariance matrix Σ. With a slight abuse of notation, we sometimes use $\mathcal{N}_d(z; \mu, \Sigma)$ to denote the corresponding density evaluated at z. The symbol $\|\cdot\|$ refers to the L_2 norm. Let $\mathbb{I}[\cdot]$ denote the standard 0-1 indicator function. Let $\Phi(\cdot)$ denote the standard Gaussian CDF, and let z_α be its upper $1-\alpha$ quantile. For a square matrix A, let $\mathrm{diag}(A)$ denote the diagonal matrix formed by zeroing out the off-diagonal entries of A, and let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimum and the maximum eigenvalues of A. We write the identity matrix as I. For sequences of constants a_n and b_n, we write $a_n = O(b_n)$ if there exists a universal constant C such that $|a_n/b_n| \leq C$ for all n larger than some n_0, and we write $a_n = o(b_n)$ if $a_n/b_n \to 0$. Similarly, for a sequence of random variables X_n and a corresponding set of constants a_n, we write $X_n = O_P(a_n)$ if $a_n^{-1} X_n$ is stochastically bounded and $X_n = o_P(a_n)$ if $a_n^{-1} X_n$ converges to zero in probability.

7.2 Background

In this section, we introduce the two main topics that we study in this chapter: two-sample mean testing using Hotelling-type statistics, and Fisher's linear discriminant analysis (LDA). We only introduce the basic versions here, later introducing high-dimensional variants like naive Bayes. In both these problems, we will

be working in the high-dimensional Gaussian setting under which the number of samples n and dimension d can both increase to infinity simultaneously.

7.2.1 Two-sample mean testing

Suppose that $X_1, \ldots, X_{n_0}, Y_1, \ldots, Y_{n_1}$ are independent random vectors in $\mathbb{R}^d$ such that $\mathcal{X}_1^{n_0} \stackrel{\text{def}}{=} \{X_1, \ldots, X_{n_0}\}$ are identically distributed with the distribution $\mathbb{P}_0$ and $\mathcal{Y}_1^{n_1} \stackrel{\text{def}}{=} \{Y_1, \ldots, Y_{n_1}\}$ are identically distributed with the distribution $\mathbb{P}_1$. Given these samples, the two-sample problem aims at testing whether

$$H_0 : \mathbb{P}_0 = \mathbb{P}_1 \quad \text{vs.} \quad H_1 : \mathbb{P}_0 \neq \mathbb{P}_1. \tag{7.1}$$

The focus of this chapter is on the specific case where $\mathbb{P}_0$ and $\mathbb{P}_1$ are d-variate Gaussian distributions with densities $p_0(x) \stackrel{\text{def}}{=} \mathcal{N}_d(x; \mu_0, \Sigma)$ and $p_1(y) \stackrel{\text{def}}{=} \mathcal{N}_d(y; \mu_1, \Sigma)$, respectively, while we discuss the extension to elliptical distributions in Section 7.8. Since we assume that $\mathbb{P}_0$ and $\mathbb{P}_1$ are Gaussians with equal covariance, the previous problem boils down to testing whether two distributions have the same mean vector or not. This two-sample mean testing is a fundamental decision-theoretic problem, having a long history in statistics; for example, the past century has seen a wide adoption of the T^2-statistic by Hotelling (1931) to decide if two-samples have different population means (Hu and Bai, 2016, for a review). Given the sample mean vectors $\widehat{\mu}_0 \stackrel{\text{def}}{=} \sum_{i=1}^{n_0} X_i/n_0$ and $\widehat{\mu}_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n_1} Y_i/n_1$ and the pooled sample covariance matrix

$$\widehat{\Sigma} \stackrel{\text{def}}{=} \frac{1}{n_0 + n_1 - 2} \left[\sum_{i=1}^{n_0} (X_i - \widehat{\mu}_0)(X_i - \widehat{\mu}_0)^\top + \sum_{i=1}^{n_1} (Y_i - \widehat{\mu}_1)(Y_i - \widehat{\mu}_1)^\top \right],$$

Hotelling's T^2-statistic is given by

$$T_H = (\widehat{\mu}_0 - \widehat{\mu}_1)^\top \widehat{\Sigma}^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1).$$

Hotelling's T^2 test based on T_H was introduced in the parametric setting for Gaussians, but it has been generalized to multivariate non-Gaussian settings as well (e.g., Kariya, 1981).

7.2.2 Fisher's linear discriminant classifier

Consider the same distributional setting described in the previous section. Given the samples $\mathcal{X}_1^{n_0}$ and $\mathcal{Y}_1^{n_1}$, classification is the problem of predicting to which of classes a new observation Z belongs, i.e. we want to predict whether Z came from $\mathbb{P}_0$ or $\mathbb{P}_1$.

Let the samples from $\mathbb{P}_0$ and $\mathbb{P}_1$ be given labels 0 and 1, respectively. If μ_0, μ_1 and Σ are known, then the optimal classifier under the Gaussian setting is given by Bayes rule:

$$\mathbb{I}\left[\log \frac{p_1(Z)}{p_0(Z)} > 0\right] = \mathbb{I}\left[(\mu_1 - \mu_0)^\top \Sigma^{-1}\left(Z - \frac{(\mu_0 + \mu_1)}{2}\right) > 0\right].$$

We denote $\delta \stackrel{\text{def}}{=} \mu_1 - \mu_0$ and $\mu_{\text{pool}} \stackrel{\text{def}}{=} (\mu_0 + \mu_1)/2$ so that we can succinctly write the Bayes rule as

$$C_{\text{Bayes}}(Z) \stackrel{\text{def}}{=} \mathbb{I}\Big[\delta^\top \Sigma^{-1}(Z - \mu_{\text{pool}}) > 0\Big]. \tag{7.2}$$

Then, by plugging in the empirical estimators $\widehat{\delta} \stackrel{\text{def}}{=} \widehat{\mu}_1 - \widehat{\mu}_0$ and $\widehat{\mu}_{\text{pool}} \stackrel{\text{def}}{=} (\widehat{\mu}_0 + \widehat{\mu}_1)/2$, Linear Discriminant Analysis (LDA) classification rule is given by

$$\text{LDA}_{n_0,n_1}(Z) \stackrel{\text{def}}{=} \mathbb{I}\Big[\widehat{\delta}^\top \widehat{\Sigma}^{-1}(Z - \widehat{\mu}_{\text{pool}}) > 0\Big].$$

The same classifier was derived by Fisher (1936, 1940) from a generalized eigenvalue problem (hence also called Fisher's LDA) and was later developed further by Wald (1944) and Anderson (1951). Define the error of LDA conditioned on the input data as:

$$\begin{aligned} \mathcal{E} &\stackrel{\text{def}}{=} (\mathcal{E}_0 + \mathcal{E}_1)/2, \\ \text{where} \quad \mathcal{E}_0 &\stackrel{\text{def}}{=} \Pr_{Z\sim\mathbb{P}_0}(\text{LDA}_{n_0,n_1}(Z) = 1 \mid \mathcal{X}_1^{n_0}, \mathcal{Y}_1^{n_1}), \\ \mathcal{E}_1 &\stackrel{\text{def}}{=} \Pr_{Z\sim\mathbb{P}_1}(\text{LDA}_{n_0,n_1}(Z) = 0 \mid \mathcal{X}_1^{n_0}, \mathcal{Y}_1^{n_1}). \end{aligned} \tag{7.3}$$

Clearly, $\mathcal{E}$ is a random variable that depends on the input data. Next, define the *unconditional* error of LDA as

$$\begin{aligned} E &\stackrel{\text{def}}{=} (E_0 + E_1)/2, \\ \text{where} \quad E_0 &\stackrel{\text{def}}{=} \mathbb{E}_{n_0,n_1}\left[\Pr_{Z\sim\mathbb{P}_0}(\text{LDA}_{n_0,n_1}(Z) = 1 \mid \mathcal{X}_1^{n_0}, \mathcal{Y}_1^{n_1})\right], \\ E_1 &\stackrel{\text{def}}{=} \mathbb{E}_{n_0,n_1}\left[\Pr_{Z\sim\mathbb{P}_1}(\text{LDA}_{n_0,n_1}(Z) = 0 \mid \mathcal{X}_1^{n_0}, \mathcal{Y}_1^{n_1})\right], \end{aligned} \tag{7.4}$$

where $\mathbb{E}_{n_0,n_1}$ denotes the expectation with respect to the n_0 and n_1 input points from each class. Note that since the input data has already been integrated out, E, E_0, E_1 do not depend on the input data and are only functions of $d, \delta, \Sigma, n \stackrel{\text{def}}{=} n_0 + n_1$.

However, E is unknown, but one can estimate E in a few different ways. One simple way is via sample splitting where the samples are split into training and test sets. Let us denote the number of samples of each class in the training set by $n_{0,\text{tr}}$ and $n_{1,\text{tr}}$. Similarly let us write the number of samples of each class in

the test set by $n_{0,\mathrm{te}}$ and $n_{1,\mathrm{te}}$. In other words, there are $n_{\mathrm{tr}} \stackrel{\mathrm{def}}{=} n_{0,\mathrm{tr}} + n_{1,\mathrm{tr}}$ samples in the training set and $n_{\mathrm{te}} \stackrel{\mathrm{def}}{=} n_{0,\mathrm{te}} + n_{1,\mathrm{te}}$ samples in the test set. We then form the LDA classifier using n_{tr} samples in the training set, and estimate its test error using the remaining n_{te} samples in the test set. Specifically, let $\mathrm{LDA}_{n_{0,\mathrm{tr}},n_{1,\mathrm{tr}}}$ be the LDA classifier formed based on the training set. Then the sample-splitting error, denoted by $\widehat{E}^S$, is given as

$$\begin{aligned} \widehat{E}^S &\stackrel{\mathrm{def}}{=} (\widehat{E}_0^S + \widehat{E}_1^S)/2, \qquad (7.5) \\ \text{where} \quad \widehat{E}_0^S &\stackrel{\mathrm{def}}{=} \frac{1}{n_{0,\mathrm{te}}} \sum_{i=1}^{n_{0,\mathrm{te}}} \mathbb{I}\Big[\mathrm{LDA}_{n_{0,\mathrm{tr}},n_{1,\mathrm{tr}}}(X_{n_{0,\mathrm{tr}}+i}) = 1\Big], \\ \widehat{E}_1^S &\stackrel{\mathrm{def}}{=} \frac{1}{n_{1,\mathrm{te}}} \sum_{i=1}^{n_{1,\mathrm{te}}} \mathbb{I}\Big[\mathrm{LDA}_{n_{0,\mathrm{tr}},n_{1,\mathrm{tr}}}(Y_{n_{1,\mathrm{tr}}+i}) = 0\Big]. \end{aligned}$$

It is clear from the definitions that the LDA classifier will have a true accuracy significantly above half if and only if $\mu_0 \neq \mu_1$. This implies that one can actually use $\widehat{E}^S$ as a test statistic for two-sample testing, by checking whether $\widehat{E}^S$ is significantly different from half or not. We shall derive the power of such a test in Section 7.6 and compare it to the best possible power (in a minimax sense).

7.3 Lower bounds for two-sample mean testing

Before we present our analysis on the power of two-sample testing via classification, we begin by understanding the fundamental minimax lower bounds for two-sample testing.

We first introduce some notation. Let $\mathcal{P}$ be a set that consists of all pairs of d-dimensional multivariate normal density functions whose covariance matrices coincide, and is positive definite. Let $\mathcal{P}_0$ be the subset of $\mathcal{P}$ such that each pair also has the same mean. For a given $\alpha \in (0,1)$, let us write a level α test based on $\mathcal{X}_1^{n_0}$ and $\mathcal{Y}_1^{n_1}$ by φ_α and the collection of all level α tests by

$$\mathcal{T}_\alpha \stackrel{\mathrm{def}}{=} \big\{\varphi_\alpha : \mathcal{X}_1^{n_0} \cup \mathcal{Y}_1^{n_1} \mapsto \{0,1\} : \sup_{p_0,p_1 \in \mathcal{P}_0} \mathbb{E}_{p_0,p_1}[\varphi_\alpha] \leq \alpha\big\}.$$

Additionally, we define a class of two multivariate normal density functions p_0 and p_1 whose distance is measured in terms of Mahalanobis distance parameterized by $\rho > 0$ as:

$$\mathcal{P}_1(\rho) \stackrel{\mathrm{def}}{=} \{(p_0,p_1) \in \mathcal{P} : (\mu_0 - \mu_1)^\top \Sigma^{-1} (\mu_0 - \mu_1) \geq \rho^2\}.$$

The use of Mahalanobis distance is conventional and has been considered in Giri et al. (1963), Giri and Kiefer (1964) and Salaevskii (1971) to study the minimax character of Hotelling's one-sample test. The "oracle"

Hotelling's two sample test is defined as

$$\varphi_H^* = \mathbb{I}\left[\frac{n_0 n_1}{n_0+n_1}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \Sigma^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) \geq c_{\alpha,d}\right],$$

where $c_{\alpha,d}$ is the $1-\alpha$ quantile of the chi-squared distribution with d degrees of freedom, and "oracle" signifies that Σ is known. Luschgy (1982) extends the previous one-sample results and shows that φ_H^* is minimax optimal over $\mathcal{P}_1(\rho)$, or more explicitly,

$$\sup_{\varphi_\alpha \in \mathcal{T}_\alpha} \inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_\alpha] = \inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_H^*], \tag{7.6}$$

for any finite n and d. However, this result does not clearly show how the underlying parameters (e.g., n, d, ρ) interact to determine the power. To shed light on this, we study the asymptotic expression for the minimax power. Let us denote the sample size ratio by $\lambda_1 = \lambda_{1,n} \stackrel{\text{def}}{=} n_1/n$. Recalling that Φ is the standard normal CDF and z_α its $1-\alpha$ quantile, we prove the following:

Proposition 7.1. *Consider a high-dimensional regime where $n, d \to \infty$ (at any rate). Then the minimax power for Gaussian two-sample mean testing is*

$$\sup_{\varphi_\alpha \in \mathcal{T}_\alpha} \inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_\alpha] = \Phi\left(-\frac{\sqrt{2d}}{\sqrt{2d + n\lambda_1(1-\lambda_1)\rho^2}} z_\alpha + \frac{n\lambda_1(1-\lambda_1)\rho^2}{\sqrt{2d + 4n\lambda_1(1-\lambda_1)\rho^2}}\right) + o(1). \tag{7.7}$$

The proof of the above result is based on the central limit theorem and can be found in Appendix F.3.2. Notably, the expression (7.7) is asymptotically precise including all constant terms and is valid without any restrictions on d/n and λ_1. The way to interpret the bound in (7.7) is as follows. The first term inside the parentheses is not of interest for our purposes, its magnitude being bounded by the constant z_α. The second term is what determines the rate at which the power approaches one. When $\rho = 0$, the power reduces to $\Phi(-z_\alpha) = \alpha$ and if d and n are thought of as fixed, larger ρ leads to larger power. The key in high dimensions, however, is how the power depends jointly on the signal to noise ratio (SNR) ρ, the dimension d and the sample size n. To see this clearer, in the low SNR regime where $\rho^2 = o(d/n)$ and $\lambda_1 \to \lambda \in (0,1)$, the minimax lower bound simplifies to

$$\Phi\left(-z_\alpha + \frac{n\lambda(1-\lambda)\rho^2}{\sqrt{2d}}\right) + o(1). \tag{7.8}$$

It can be already seen that at constant SNR, n only needs to scale faster than $\sqrt{d}$ for test power to asymptotically approach unity — this $\sqrt{d}/n$ scaling is unlike the d/n scaling that one typically sees in prediction problems (for prediction error or classifier recovery, see e.g. Raudys and Young, 2004). In the next section, we prove that this lower bound is tight even when Σ is unknown, as long as $d = o(n)$.

7.4 Minimax optimality of Hotelling's test when $d = o(n)$

When Σ is unknown, φ_H^* is not available and thus it remains unclear whether the previous asymptotic lower bound is tight. In other words, we do not know whether there exists a test that has the same asymptotic minimax power as φ_H^* in all high-dimensional regimes with unknown Σ. In this section, we will make a first step towards closing this gap. In particular, we will show that Hotelling's test with unknown Σ can achieve the same asymptotic minimax power as φ_H^* when $d/n \to 0$. By letting $q_{\alpha,n,d}$ be the $1-\alpha$ quantile of the F distribution with parameters d and $n-1-d$, Hotelling's two-sample test with unknown Σ is given by

$$\varphi_H = \mathbb{I}\left[\frac{n_0 n_1 (n-d-1)}{n(n-2)d}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \widehat{\Sigma}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) \geq q_{\alpha,n,d}\right].$$

For Gaussians, it is well-known that φ_H satisfies $\sup_{p_0,p_1\in\mathcal{P}_0} \mathbb{E}_{p_0,p_1}[\varphi_H] \leq \alpha$ (e.g., Anderson, 2003). The next theorem studies the power of φ_H.

Theorem 7.1. *Consider an asymptotic regime where $d/n \to 0$. Then the uniform power of φ_H is asymptotically the same as that of φ_H^* for Gaussian two-sample mean testing. In other words, as $n, d \to \infty$ with $d/n \to 0$, we have that $\inf_{p_0,p_1\in\mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_H]$ is equal to*

$$\Phi\left(-\frac{\sqrt{2d}}{\sqrt{2d + n\lambda_1(1-\lambda_1)\rho^2}} z_\alpha + \frac{n\lambda_1(1-\lambda_1)\rho^2}{\sqrt{2d + 4n\lambda_1(1-\lambda_1)\rho^2}}\right) + o(1).$$

The proof of the above theorem can be found in Appendix F.3.3. Theorem 7.1 is in contrast to previous negative results on the high-dimensional behavior of Hotelling's test. For example, Bai and Saranadasa (1996) demonstrate that the performance of φ_H can be bad when $d/n \to c \in (0,1)$. When the dimension is larger than the sample size, Hotelling's test statistic T_H is not even well-defined. Due to its limitations, Hotelling's test has been largely neglected in the setting where d increases with n. Unlike the previous negative results, Theorem 7.1 revives φ_H by showing that it achieves the minimax power under the asymptotic regime where d is allowed to grow, but $d/n \to 0$. We also provide empirical support for our asymptotic results in Figure 7.4 of Section 7.10.3.

Remark 7.1. *Combining the previous theorem with Bai and Saranadasa (1996) and our simulation results in Section 7.10.3, we may describe the phase transition behavior of Hotelling's test with unknown Σ as*

- *Optimal regime (same power as φ_H^*): $d/n \to 0$,*
- *Suboptimal regime (lower power than φ_H^*): $d/n \to c \in (0,1)$,*
- *Not applicable: $d/n \to c \geq 1$.*

Even though Hotelling's test is suboptimal when $d = O(n)$, it is still an open problem to determine whether the lower bound is achievable by some other test, or whether a stronger lower bound can be proved.

7.5 Asymptotic normality of the accuracy of generalized LDA

Here, we investigate the high-dimensional limiting distribution of the sample-splitting error in (7.5). Building on the results developed in this section, we will present the power of the classification test in Section 7.6. Our main interest is in the setting where the dimension is comparable to or potentially much larger than the sample size. In this high-dimensional scenario, Bickel and Levina (2004) prove that Fisher's LDA performs poorly in classification problems. When $d > n$, Fisher's LDA classifier is not even well-defined since $\widehat{\Sigma}$ is not invertible. Thus, Bickel and Levina (2004) consider the naive Bayes (NB) classification rule by replacing $\widehat{\Sigma}^{-1}$ with the inverse of $\mathrm{diag}(\widehat{\Sigma})$ and show that it outperforms Fisher's LDA in the high-dimensional setting. In the context of two-sample testing, we encounter the same issue on $\widehat{\Sigma}$ as mentioned earlier. To simplify our analysis, we start by assuming that Σ is *known* and analyze the asymptotic behavior of the corresponding Fisher's LDA statistic. Later in Section 7.7, we extend the results to *unknown* Σ by considering the NB classifier and others.

7.5.1 Assumptions

Recalling that we work in the high-dimensional Gaussian setting with common covariance, let us detail some assumptions that facilitate our analysis. We assume that as $n = n_0 + n_1 \to \infty$, we have

(A1) *High-dimensional asymptotics*: there exists $c \in (0, \infty)$ such that $d/n \to c$.

(A2) *Local alternative*: $\delta^\top \Sigma^{-1} \delta = O(n^{-1/2})$.

(A3) *Sample size ratio*: there exists $\lambda \in (0, 1)$ such that $n_0/n \to \lambda$.

(A4) *Sample splitting ratio*: there exists $\kappa \in (0, 1)$ such that $n_{\mathrm{tr}}/n \to \kappa$.

The asymptotic regime in **(A1)** is called *Raudys-Kolmogorov double asymptotics* (e.g. Zollanvari et al., 2011) and assumes that d increases linearly with n. In **(A2)**, we assume that $\delta^\top \Sigma^{-1} \delta$ is close to zero such that a minimax test has nontrivial power. Note that under **(A1)**, the low SNR regime $\delta^\top \Sigma^{-1} \delta = o(d/n)$ is implied by **(A2)**. It is also interesting to note that the classification error of the Bayes optimal classifier (7.2) is computed as

$$\frac{1}{2} \Pr_{Z \sim \mathbb{P}_0} \left\{ C_{\mathrm{Bayes}}(Z) = 1 \right\} + \frac{1}{2} \Pr_{Z \sim \mathbb{P}_1} \left\{ C_{\mathrm{Bayes}}(Z) = 0 \right\} = 1 - \Phi\left(\frac{\sqrt{\delta^\top \Sigma^{-1} \delta}}{2} \right),$$

which means that the classification error of the Bayes classifier, and hence *any* classifier, approaches chance under **(A2)**. Assumption **(A3)** rules out highly imbalanced cases and is common in the two-sample literature (e.g. Bai and Saranadasa, 1996; Chen and Qin, 2010; Srivastava et al., 2013). **(A4)** assumes that the user-chosen sample-splitting ratio is within $(0, 1)$. We show in Theorem 7.3 that the asymptotic power of the test based on held-out classification accuracy is maximized when $\kappa = 1/2$ for the balanced case of $\lambda = 1/2$. In other cases, Theorem 7.3 may serve as a guideline for choosing κ that maximizes the asymptotic power. For any $d \times d$ symmetric positive definite matrix A, we define the generalized LDA classifier by

$$\mathrm{LDA}_{A,n_0,n_1}(Z) \stackrel{\mathrm{def}}{=} \mathbb{I}\Big[\widehat{\delta}^\top A(Z - \widehat{\mu}_{\mathrm{pool}}) > 0\Big]. \tag{7.9}$$

Its error can be calculated by replacing $\mathrm{LDA}_{n_{0,\mathrm{tr}},n_{1,\mathrm{tr}}}(Z)$ with $\mathrm{LDA}_{A,n_{0,\mathrm{tr}},n_{1,\mathrm{tr}}}(Z)$ in expression (7.5):

$$\widehat{E}^S_A \equiv \text{classification error of } \mathrm{LDA}_{A,n_{0,\mathrm{tr}},n_{1,\mathrm{tr}}}(Z),$$

emphasizing the dependency on the user-chosen matrix A. In terms of Σ and A, we assume that:

(A5) Σ *has bounded eigenvalues:* there exist constants c_1, c_2 such that $0 < c_1 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le c_2 < \infty$.

(A6) A *has bounded eigenvalues:* there exist constants c_1', c_2' such that $0 < c_1' \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le c_2' < \infty$.

The same eigenvalue condition for Σ was used by Bickel and Levina (2004). Assumption **(A6)** is satisfied when A is diagonal with uniformly bounded entries, and when $A = \Sigma^{-1}$ under **(A5)**.

7.5.2 Asymptotic normality for non-random A

Given the previous assumptions, we study the asymptotic distribution of the sample-splitting error of the generalized LDA classifier when A is non-random. Since Fisher's LDA with known Σ is a special case of the generalized LDA classifier, it is straightforward to derive the limiting distribution of $\widehat{E}^S_{\Sigma^{-1}}$ from the general result.

We first observe that the sample-splitting error of the generalized LDA classifier can be viewed as the average of independent observations when conditioning on the training set. Therefore it is natural to expect that the sample-splitting error is asymptotically normally distributed. To make this statement formal, we define $\mathcal{E}_{i,A}$ and $E_{i,A}$ similarly as $\mathcal{E}_i$ and E_i for $i = 1, 2$ from definitions (7.3) and (7.4), but by replacing the LDA classifier with the generalized LDA classifier with a given A. Then let us write the standardized test

statistic as

$$W_A \stackrel{\text{def}}{=} \frac{\widehat{E}_A^S - \mathcal{E}_{0,A}/2 - \mathcal{E}_{1,A}/2}{\sqrt{\mathcal{E}_{0,A}(1-\mathcal{E}_{0,A})/(4n_{0,\text{te}}) + \mathcal{E}_{1,A}(1-\mathcal{E}_{1,A})/(4n_{1,\text{te}})}}. \tag{7.10}$$

In the next proposition, we present both *conditional* and *unconditional* limiting distributions of W_A in the high dimensional setting.

Proposition 7.2. *Suppose that the assumptions* ***(A1)****–****(A6)*** *hold. Then W_A converges to the standard normal distribution conditional on the training set:*

$$\sup_{t\in\mathbb{R}} |\Pr(W_A \le t | \mathcal{X}_1^{n_{0,\text{tr}}}, \mathcal{Y}_1^{n_{1,\text{tr}}}) - \Phi(t)| = O_P(n^{-1/2}).$$

Moreover, under the same assumptions, W_A converges to the standard normal distribution unconditional on the training set:

$$\sup_{t\in\mathbb{R}} |\Pr(W_A \le t) - \Phi(t)| = o(1).$$

The proof of Proposition 7.2 can be found in Appendix F.3.4. Although the limiting distribution of W_A is known from the previous lemma, it is quite challenging to determine the power of a test based classification accuracy by analyzing W_A. The reason is that $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ are random since they depend on the training set. To address this issue, we shall present a tractable approximation of W_A that replaces $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ with non-random quantities. To ease notation, let us denote $V_{0,A} \stackrel{\text{def}}{=} \widehat{\delta}^\top A(\mu_0 - \widehat{\mu}_{\text{pool}})$, $V_{1,A} \stackrel{\text{def}}{=} \widehat{\delta}^\top A(\widehat{\mu}_{\text{pool}} - \mu_1)$ and $U_A \stackrel{\text{def}}{=} \widehat{\delta}^\top A\Sigma A\widehat{\delta}$. We would like to stress that $\widehat{\delta}$ and $\widehat{\mu}_{\text{pool}}$ are computed based only on the training set. Using this fact, $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ can be written as

$$\mathcal{E}_{0,A} = \Phi\left(\frac{V_{0,A}}{\sqrt{U_A}}\right) \quad \text{and} \quad \mathcal{E}_{1,A} = \Phi\left(\frac{V_{1,A}}{\sqrt{U_A}}\right). \tag{7.11}$$

Further write the expectations of $V_{0,A}$, $V_{1,A}$ and U_A by $\mathbb{E}[V_{0,A}] = \Psi_{A,n,d} + \Xi_{A,n,d}$, $\mathbb{E}[V_{1,A}] = \Psi_{A,n,d} - \Xi_{A,n,d}$ and $\mathbb{E}[U_A] = \Lambda_{A,n,d}$ where

$$\begin{aligned} \Psi_{A,n,d} &\stackrel{\text{def}}{=} -\frac{1}{2}\delta^\top A\delta, \\ \Lambda_{A,n,d} &\stackrel{\text{def}}{=} \delta^\top A\Sigma A\delta + \left(\frac{1}{n_{0,\text{tr}}} + \frac{1}{n_{1,\text{tr}}}\right)\text{tr}\big\{(A\Sigma)^2\big\}, \\ \text{and} \quad \Xi_{A,n,d} &\stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{1}{n_{0,\text{tr}}} - \frac{1}{n_{1,\text{tr}}}\right)\text{tr}(A\Sigma). \end{aligned} \tag{7.12}$$

Here the first two terms $\Psi_{A,n,d}$ and $\Lambda_{A,n,d}$ can be viewed as signal and noise terms, respectively, which ultimately determine the asymptotic power of the accuracy test. The third term $\Xi_{A,n,d}$ is an extra variance that comes from unbalanced sample sizes. Finally, we define a scaling factor

$$\gamma_{A,n,d} \stackrel{\text{def}}{=} 2\sqrt{\frac{n_{0,\text{te}}n_{1,\text{te}}}{n_{0,\text{te}}+n_{1,\text{te}}}}\frac{1}{\sqrt{\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\{1-\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\}}}. \tag{7.13}$$

With this notation in hand and letting $\phi(\cdot)$ be the standard normal density function, we now introduce an approximation of W_A defined as

$$W_A^\dagger \stackrel{\text{def}}{=} \gamma_{A,n,d}\cdot\left\{\widehat{E}_A^S-\frac{1}{2}-\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right\}.$$

It is clear that $W_A^\dagger$ is centered and scaled by explicit and non-random quantities. The next theorem shows that the difference between W_A and $W_A^\dagger$ is asymptotically negligible and therefore $W_A^\dagger$ is also asymptotically standard normal.

Theorem 7.2. *Suppose that the assumptions **(A1)**–**(A6)** hold. Then we have that $W_A = W_A^\dagger + o_P(1)$ and thus the distribution of $W_A^\dagger$ converges to a standard normal:*

$$\sup_{t\in\mathbb{R}}|\Pr(W_A^\dagger \le t)-\Phi(t)| = o(1).$$

The proof of Theorem 7.2 can be found in Appendix F.3.5. The asymptotic normality, established in the above theorem, holds under the null as well as under the local alternative **(A2)**. This enables us to explore the asymptotic power of the generalized LDA test with known Σ in the next section, and we deal with unknown Σ in the following section.

7.6 Asymptotic power of generalized LDA with non-random A

Here, we study the asymptotic power of the generalized LDA test for known Σ. Since a smaller value of $\widehat{E}_A^S - 1/2$ (or equivalently a larger value of the average per-class accuracy $1-\widehat{E}_A^S$) is in favor of $H_1 : \mu_0 \neq \mu_1$, we define the test function by

$$\varphi_A \stackrel{\text{def}}{=} \mathbb{I}\left[\gamma_{A,n,d}\left(\widehat{E}_A^S-\frac{1}{2}\right) < -z_\alpha\right]. \tag{7.14}$$

It is then clear from Theorem 7.2 that φ_A has an asymptotic type-1 error controlled by α. Now under the local alternative hypothesis, φ_A has power given by

$$\begin{aligned}\mathbb{E}[\varphi_A] &= \Pr\left(W_A^\dagger < -z_\alpha - \gamma_{A,n,d}\cdot\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right),\\ &= \Phi\left(-z_\alpha - \gamma_{A,n,d}\cdot\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + o(1),\end{aligned} \tag{7.15}$$

where the second equality uses Theorem 7.2. Let us write

$$\beta_{A,\lambda,\kappa} \stackrel{\text{def}}{=} \frac{\lambda - 1/2}{\sqrt{\lambda(1-\lambda)\kappa}}\frac{n^{-1}\mathrm{tr}(A\Sigma)}{\sqrt{n^{-1}\mathrm{tr}\{(A\Sigma)^2\}}}. \tag{7.16}$$

Using assumptions **(A1)**–**(A6)**, the main term in the power function (7.15) simplifies as

$$-\gamma_{A,n,d}\cdot\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}} = \frac{\sqrt{2\kappa(1-\kappa)}\phi(\beta_{A,\lambda,\kappa})}{\sqrt{\Phi(\beta_{A,\lambda,\kappa})\{1-\Phi(\beta_{A,\lambda,\kappa})\}}}\cdot\frac{n\lambda(1-\lambda)\delta^\top A\delta}{\sqrt{2\mathrm{tr}\{(A\Sigma)^2\}}} + o(1).$$

Resubstituting the above into expression (7.15), we finally infer that

$$\mathbb{E}[\varphi_A] = \Phi\left(-z_\alpha + \frac{\sqrt{2\kappa(1-\kappa)}\phi(\beta_{A,\lambda,\kappa})}{\sqrt{\Phi(\beta_{A,\lambda,\kappa})\{1-\Phi(\beta_{A,\lambda,\kappa})\}}}\cdot\frac{n\lambda(1-\lambda)\delta^\top A\delta}{\sqrt{2\mathrm{tr}\{(A\Sigma)^2\}}}\right) + o(1). \tag{7.17}$$

Since $\sup_{x\in\mathbb{R}}\phi(x)/\sqrt{\Phi(x)\{1-\Phi(x)\}} = \sqrt{2/\pi}$ and its maximum is achieved at $x = 0$, the asymptotic power (7.17) is maximized when $\lambda = 1/2$ and $\kappa = 1/2$, further supported by simulations in Appendix F.4. However it is unknown whether the same result continues to hold for a random A (e.g. $A = \widehat{\Sigma}^{-1}$). In this balanced setting, the asymptotic power is further simplified as

$$\Phi\left(-z_\alpha + \frac{n\delta^\top A\delta}{\sqrt{32\pi\mathrm{tr}\{(A\Sigma)^2\}}}\right) + o(1). \tag{7.18}$$

For ease of reference, we summarize our discussion as a theorem.

Theorem 7.3. *Suppose that the assumptions **(A1)**–**(A6)** hold. Then the generalized LDA test (7.14) asymptotically controls type-1 error at level α and its power for Gaussian two-sample mean testing is given by*

$$\mathbb{E}[\varphi_A] = \Phi\left(-z_\alpha + \frac{\sqrt{2\kappa(1-\kappa)}\phi(\beta_{A,\lambda,\kappa})}{\sqrt{\Phi(\beta_{A,\lambda,\kappa})\{1-\Phi(\beta_{A,\lambda,\kappa})\}}}\cdot\frac{n\lambda(1-\lambda)\delta^\top A\delta}{\sqrt{2\mathrm{tr}\{(A\Sigma)^2\}}}\right) + o(1). \tag{7.19}$$

Furthermore, keeping other parameters fixed, the asymptotic power is maximized when $\lambda = 1/2$ and $\kappa = 1/2$ (corresponding to a balanced train/test split).

The proof of the above theorem follows immediately from the previous discussion and so is omitted. As a direct consequence of Theorem 7.3, when $\lambda = 1/2$ and $\kappa = 1/2$, the power of the "oracle" Fisher's LDA test that uses $A = \Sigma^{-1}$ (again, "oracle" is used because it uses Σ^{-1}) becomes

$$\mathbb{E}[\varphi^*_{\Sigma^{-1}}] = \Phi\bigg(- z_\alpha + \frac{n\delta^\top \Sigma^{-1}\delta}{\sqrt{32\pi d}} \bigg) + o(1). \tag{7.20}$$

Comparing the above power with the minimax lower bound expression (7.8) with $\lambda = 1/2$, we may conclude that the classification accuracy test can achieve essentially minimax optimal power, up to the small constant factor $1/\sqrt{\pi} \approx 0.564$. In other words, we pay a constant factor by performing a two-sample testing via classification compared to the minimax optimal test. However, this somewhat positive conclusion should be treated with caution as emphasized below:

- First, Theorem 7.3 is a pointwise result. That means, the result holds for any sequence of distributions satisfying the assumptions, but not uniformly over a class of distributions. Hence, conceptually, this is weaker than the uniform power achieved by φ^*_H in Theorem 7.1. However, this drawback actually applies to almost every published result on high-dimensional two-sample testing that we are aware of (or certainly all those that we cite), and it is a much broader open problem to prove that the power guarantees for these tests hold uniformly over the relevant classes.

- Second, although a constant factor is not of major concern in determining the minimax rate, it may have a significant effect on power in practice. To see this, let n_{Fisher} and $n_{\text{Hotelling}}$ be the sample sizes needed for $\varphi^*_{\Sigma^{-1}}$ and φ^*_H to obtain the same power against the local alternative considered in Theorem 7.3. Then the asymptotic relative efficiency (ARE) of $\varphi^*_{\Sigma^{-1}}$ with respect to φ^*_H is defined as the limit of the ratio $n_{\text{Hotelling}}/n_{\text{Fisher}}$ (e.g. Chapter 14 of Van der Vaart, 2000). Based on the asymptotic power expressions (7.8) and (7.19), a simple closed-form expression of the ARE is available as

$$\text{ARE}(\varphi^*_{\Sigma^{-1}}; \varphi^*_H) = \frac{\sqrt{2\kappa(1-\kappa)}\phi(\beta^*)}{\sqrt{\Phi(\beta^*)\{1-\Phi(\beta^*)\}}} \leq \frac{1}{\sqrt{\pi}} \approx 0.564, \tag{7.21}$$

  where $\beta^* = \lim_{n,d\to\infty} \beta_{\Sigma^{-1},\lambda,\kappa}$ if it exists. This ARE expression implies that $\varphi^*_{\Sigma^{-1}}$ requires (at least) $\sqrt{\pi} \approx 1.77$ more samples to attain approximately the same power as φ^*_H. In this context, Hotelling's test should be preferred over the classifier-based test to obtain higher power against the Gaussian mean shift alternative.

In the following sections, we extend the results on the oracle Fisher's LDA classifier to it variants with unknown Σ and also to elliptical distributions.

Remark 7.2. As mentioned in Section 7.5.1, the accuracy of the Bayes optimal classifier approaches half under the considered asymptotic regime, meaning that no classifier can have accuracy better than a random guess in the limit. In contrast, under the same asymptotic regime, two-sample testing based on generalized LDA can have non-trivial power (strictly greater than α) as shown in Theorem 7.3. These two results not only demonstrate that testing is easier than classification, but also that the local alternative **(A2)** is conceptually interesting — it corresponds to a regime where the LDA classifier performs as poorly as a random guess for classification, but is essentially optimal for testing.

7.7 Naive Bayes: power of generalized LDA with unknown Σ

For low-dimensional Gaussians with unknown Σ, there are strong reasons to prefer Hotelling's test; it is well-known that it is *uniformly most powerful* among all tests that are invariant with respect to nonsingular linear transformations (e.g., Anderson, 2003). We also refer to Simaika (1941); Giri et al. (1963); Giri and Kiefer (1964); Salaevskii (1971); Kariya (1981); Luschgy (1982) for other optimality properties of Hotelling's test in finite d and n settings. Moreover our result in Theorem 7.1 says that φ_H is asymptotically minimax optimal among all level α tests as long as $d/n \to 0$. Unfortunately, when d is linearly comparable to or larger than n, these optimal properties of Hotelling's test becomes highly non-trivial. In particular, φ_H has asymptotic power tending to the (trivial) value of α in the high dimensional setting, when $d, n \to \infty$ with $d/n \to 1 - \epsilon$ for small $\epsilon > 0$ (Bai and Saranadasa, 1996, for details). The problem becomes even worse when the dimension is larger than the sample size as T_H is not well-defined.

The aforementioned issue on T_H has motivated the study of alternative two-sample mean test statistics in the high-dimensional setting. For instance, Bai and Saranadasa (1996) show that dropping $\widehat{\Sigma}$ from the Hotelling test statistic (i.e. replacing $\widehat{\Sigma}$ with the identity matrix) entirely leads to a test that does have asymptotic power tending to one in the high-dimensional setting where Hotelling's test fails. The test statistic proposed by Bai and Saranadasa (1996) can be essentially written as

$$T_{BS} \stackrel{\text{def}}{=} (\widehat{\mu}_0 - \widehat{\mu}_1)^\top (\widehat{\mu}_0 - \widehat{\mu}_1).$$

Following that, Srivastava and Du (2008) propose (in a similar spirit) the test statistic

$$T_{SD} \stackrel{\text{def}}{=} (\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathrm{diag}(\widehat{\Sigma})^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1), \tag{7.22}$$

by replacing $\widehat{\Sigma}$ with $\mathrm{diag}(\widehat{\Sigma})$ in Hotelling's statistic. They show that T_{SD} also leads to high-dimensional consistency.

As mentioned earlier, the idea of using $\mathrm{diag}(\widehat{\Sigma})$ in place of $\widehat{\Sigma}$ has also been justified in the high-dimensional classification problem (Bickel and Levina, 2004). In particular, the naive Bayes classifier (corresponding to T_{SD}) outperforms Fisher's LDA classifier (corresponding to T_H) in terms of the worst-case classification error in the high-dimensional setting. We note that this relatively understated connection between two-sample testing and classification has important implications for extending our previous results to other linear classifiers. Specifically, as we shall see, the power of the classifier-based tests is only worse by a constant factor than the variants of Hotelling's test when both the classifier and the two-sample test use the same substitute for Σ^{-1}.

To start, let us consider two classifiers with unknown Σ. The first one is the naive Bayes classifier and the other is the generalized LDA classifier with the identity matrix, i.e. $A = I$. We then compare the power of the corresponding classification accuracy tests with the two-sample mean tests based on T_{SD} and T_{BS}. Throughout this section, we assume that $n_0 = n_1$, $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$ and $n_{\mathrm{tr}} = n_{\mathrm{te}}$ for simplicity.

From Theorem 7.3, the asymptotic power of the test based on $\widehat{E}_I^S$ is already available as

$$\mathbb{E}[\varphi_I] = \Phi\left(-z_\alpha + \frac{n\delta^\top \delta}{\sqrt{32\pi \mathrm{tr}(\Sigma^2)}} \right) + o(1). \tag{7.23}$$

Under more general conditions than the assumptions **(A1)**–**(A6)**, Bai and Saranadasa (1996) show that the asymptotic power of the test based on T_{BS}, denoted by φ_{BS}, is

$$\mathbb{E}[\varphi_{BS}] = \Phi\left(-z_\alpha + \frac{n\delta^\top \delta}{\sqrt{32\mathrm{tr}(\Sigma^2)}} \right) + o(1). \tag{7.24}$$

Now by comparing two power expressions in (7.23) and (7.24), we arrive at the same conclusion as before that the classification accuracy test is less powerful than the corresponding two-sample test φ_{BS} by the constant factor $1/\sqrt{\pi} \approx 0.564$.

Next we focus on the naive Bayes classifier and compute the asymptotic power of the resulting test. Although the analysis proceeds similarly to the previous one, we now need to deal with the randomness from the inverse diagonal matrix, which requires extra non-trivial work. By putting $\widehat{D}^{-1} \stackrel{\mathrm{def}}{=} \mathrm{diag}(\widehat{\Sigma})^{-1}$ and $D^{-1} = \mathrm{diag}(\Sigma)^{-1}$, the asymptotic power of the naive Bayes classifier is provided as follows.

Theorem 7.4. *Consider the case where $n_0 = n_1$, $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$ and $n_{\mathrm{tr}} = n_{\mathrm{te}}$. Then under the assumptions **(A1)**, **(A2)** and **(A5)**, the power of the naive Bayes classifier test for Gaussian two-sample mean testing*

is

$$\mathbb{E}[\varphi_{\widehat{D}^{-1}}] = \Phi\bigg(- z_\alpha + \frac{n\delta^\top D^{-1}\delta}{\sqrt{32\pi \mathrm{tr}\{(D^{-1}\Sigma)^2\}}}\bigg) + o(1). \tag{7.25}$$

The proof of Theorem 7.4 can be found in Appendix F.3.5. Srivastava and Du (2008) study the asymptotic power of the test φ_{SD} based on T_{SD} (7.22). One can also check that their conditions are fulfilled under the assumptions **(A1)**–**(A5)**. Using $\lambda = 1/2$, the power of φ_{SD} is given by

$$\mathbb{E}[\varphi_{SD}] = \Phi\bigg(- z_\alpha + \frac{n\delta^\top D^{-1}\delta}{\sqrt{32 \mathrm{tr}\{(D^{-1}\Sigma)^2\}}}\bigg) + o(1).$$

Comparing this with the asymptotic power of $\varphi_{\widehat{D}^{-1}}$ in (7.25), we see that the power of the accuracy test based on the naive Bayes classifier is worse than the corresponding two-sample test φ_{SD}, once again achieving an ARE of exactly $1/\sqrt{\pi}$.

7.8 Extension to elliptical distributions

In this section we extend our main result (Theorem 7.3) to the class of elliptical distributions and show that the asymptotic power expression remains the same up to a constant factor. Let μ be a d-dimensional vector, S be a $d \times d$ positive semi-definite matrix, $\xi(\cdot)$ be a nonnegative function. A random vector Z in $\mathbb{R}^d$ is said to have an elliptical distribution with location parameter μ, scale matrix S and generator $\xi(\cdot)$ if its characteristic function satisfies

$$\mathbb{E}\big[e^{it^\top Z}\big] = e^{it^\top \mu}\xi\big(t^\top S t\big) \quad \text{for all } t \in \mathbb{R}^d.$$

When the second moment exists, it can be verified that μ corresponds to the mean vector of Z and S is proportional to the covariance matrix of Z, denoted by Σ. More specifically, by letting $\xi'(0)$ be the first derivative of ξ evaluated at zero, S is explicitly linked to Σ as $-2\xi'(0)S = \Sigma$. Notable examples of elliptical distributions include the multivariate normal, the multivariate student t, the multivariate Laplace and the multivariate logistic distribution. We refer to Gómez et al. (2003); Frahm (2004); Fang et al. (2018) for further properties and examples of elliptical distributions. To have an explicit power expression, we make two extra assumptions on Z described as follows:

(A7) *Condition on kurtosis parameter*: let ζ_{kurt} be the kurtosis parameter of Z defined as

$$\zeta_{\mathrm{kurt}} \overset{\mathrm{def}}{=} \frac{\mathbb{E}\big[\big\{(Z-\mu)^\top \Sigma^{-1}(Z-\mu)\big\}^2\big]}{d(d+2)} - 1.$$

We assume that there exists a positive constant M such that $\zeta_{\text{kurt}} < M$ for all n, d.

(A8) *Condition on density function*: assume that the standardized first coordinate of Z, that is $e_1^\top(Z-\mu)/(e_1^\top\Sigma e_1)^{1/2}$ where $e_1 = (1, 0, \ldots, 0)^\top$, has the density function $f_\xi(\cdot)$ with respect to the Lebesgue measure. We further assume that f_ξ is bounded and continuously differentiable.

We believe that the condition on ζ_{kurt} in **(A7)** is mild and satisfied for many elliptical distributions (e.g., **?**). For example, the kurtosis parameter of the multivariate t-distribution with ν degrees of freedom is $2/(\nu-4)$ for $\nu > 4$, which in turn implies that ζ_{kurt} is zero for the Gaussian case. To interpret **(A8)**, we note that each component of an elliptical random vector has the same distribution after standardization. Assumption **(A8)** then states that this common distribution has the density function f_ξ with some extra regularity conditions. Clearly f_ξ corresponds to the standard normal density function for the Gaussian case that is bounded and continuously differentiable. But **(A8)** fails to hold for the Laplace distribution whose density function is not differentiable at zero. With these extra assumptions, we are now ready to present the main result of this section, which generalizes Theorem 7.3 to elliptical distributions.

Theorem 7.5. *Suppose that $\mathbb{P}_0$ and $\mathbb{P}_1$ are elliptical distributions with parameters (μ_0, S, ξ) and (μ_1, S, ξ), respectively. Consider the case where $n_0 = n_1$, $n_{0,\text{tr}} = n_{1,\text{tr}}$ and $n_{\text{tr}} = n_{\text{te}}$, i.e. $\lambda = \kappa = 1/2$, for simplicity. Then under the assumptions **(A1)**, **(A2)** and **(A5)**–**(A8)**, the generalized LDA test (7.14) asymptotically controls type-1 error at level α and has the asymptotic power for testing the hypothesis (7.1) as*

$$\mathbb{E}[\varphi_A] = \Phi\left(-z_\alpha + \frac{f_\xi(0) \cdot n\delta^\top A\delta}{\sqrt{16\text{tr}\{(A\Sigma)^2\}}} \right) + o(1). \tag{7.26}$$

The above result shows that the asymptotic power expression in Theorem 7.3 does not change in terms of n, d, Σ, A, δ, for elliptical distributions. To further illustrate the result, let us consider the specific case where $\mathbb{P}_0$ and $\mathbb{P}_1$ are multivariate t-distributions with ν degrees of freedom and the same scale matrix. We additionally assume that $\nu > 4$ under which the assumption **(A7)** is satisfied. In such a case, $f_\xi(0) = f_\xi(0; \nu)$ can be written as

$$f_\xi(0;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)} \to \frac{1}{\sqrt{2\pi}} \approx 0.399 \quad \text{as } \nu \to \infty.$$

Hence, by taking $\nu \to \infty$, the asymptotic power (7.26) recovers the previous power expression (7.18) for the Gaussian case. Indeed $f_\xi(0;\nu)$ is a decreasing sequence of ν such that $f_\xi(0;\nu) < f_\xi(0;4) \approx 0.530$ for all $\nu > 4$. This fact demonstrates that the generalized LDA test becomes relatively more efficient when the underlying t-distributions have heavier tails, which is also validated by simulations (see Figure 7.2).

7.9 Results on general classifiers

So far we have focused on the accuracy tests based on linear classifiers and derived their explicit asymptotic power against local alternatives under Gaussian or elliptical distribution assumptions. In this section, we turn to more general settings and examine two key properties, namely the type-1 error control and consistency, of the accuracy test based on a general classifier. The main result of this section shows that a classification accuracy test achieves asymptotic power equal to one, provided that the corresponding classifier has an accuracy higher than chance. This result naturally motivates questions about rate, for which more assumptions are needed, and also motivates studying a more challenging setting where the true accuracy approaches half, like the one we consider for the generalized LDA test.

Let $\widehat{C}(\cdot)$ denote a generic classifier based on the training set which maps from $\mathcal{X}_1^{n_{0,\mathrm{tr}}} \cup \mathcal{Y}_1^{n_{1,\mathrm{tr}}}$ to $\{0,1\}$. The per-class errors of this generic classifier, denoted by $\widehat{E}_0^S(\widehat{C})$ and $\widehat{E}_1^S(\widehat{C})$, are calculated by replacing $\mathrm{LDA}_{n_{0,\mathrm{tr}},n_{1,\mathrm{tr}}}(Z)$ with $\widehat{C}(Z)$ in expression (7.5). By taking the average of these two errors, we define the sample-splitting error of $\widehat{C}$ as

$$\widehat{E}^S(\widehat{C}) \stackrel{\mathrm{def}}{=} \big\{\widehat{E}_0^S(\widehat{C}) + \widehat{E}_1^S(\widehat{C})\big\}/2, \tag{7.27}$$

and reject the null hypothesis $H_0 : \mathbb{P}_0 = \mathbb{P}_1$ if $\widehat{E}^S(\widehat{C})$ is significantly smaller than a half. To facilitate analysis, we assume the following asymptotic properties of $\widehat{E}_0^S(\widehat{C})$ and $\widehat{E}_1^S(\widehat{C})$:

(A9) *Asymptotic classification errors*: assume that $\widehat{E}_0^S(\widehat{C}) = E_0(C) + o_P(1)$ and $\widehat{E}_1^S(\widehat{C}) = E_1(C) + o_P(1)$ where $E_1(C)$ and $E_2(C)$ are constants within $(0,1)$. Moreover, there exists a strictly positive constant $\epsilon > 0$ such that $E_0(C)/2 + E_1(C)/2 = 1/2 - \epsilon$ under the alternative hypothesis.

To determine the threshold of a test, we consider two methods: (1) the Gaussian approximation that underlies our theory in the preceding sections and (2) the permutation procedure that has been common in practice with finite sample guarantees. We start by analyzing the asymptotic test based on the Gaussian approximation and then turn to permutation tests.

7.9.1 Asymptotic test

As discussed before, the sample-splitting error can be viewed as the sum of independent random variables given the training set. Therefore it is natural to expect that this empirical error follows closely a normal distribution even for a general classifier when the sample size is large. Building on this intuition, we define the asymptotic test as

$$\mathbb{I}\left[\frac{2\widehat{E}^S(\widehat{C}) - 1}{\sqrt{\widehat{E}_0^S(\widehat{C})\{1 - \widehat{E}_0^S(\widehat{C})\}/n_{0,\mathrm{te}} + \widehat{E}_1^S(\widehat{C})\{1 - \widehat{E}_1^S(\widehat{C})\}/n_{1,\mathrm{te}}}} < -z_\alpha\right]$$

and denote it by $\varphi_{\widehat{C},\text{Asymp}}$. We note that the quantity inside of the indicator function is a studentized sample-splitting error under the null hypothesis. In the next proposition we prove that the normal approximation is indeed accurate and thus $\varphi_{\widehat{C},\text{Asymp}}$ is a valid test at least asymptotically. Moreover, when the sequence of classification errors tends to a constant that is strictly less than chance level, we show that the power of the asymptotic test tends to one as $n \to \infty$ potentially with $d \to \infty$.

Proposition 7.3. *Suppose that the assumptions **(A3)**, **(A4)** and **(A9)** hold as $n \to \infty$ potentially with $d \to \infty$ at any relative rate. Then under the null hypothesis $H_0 : \mathbb{P}_0 = \mathbb{P}_1$, we have $\lim_{n\to\infty} \mathbb{E}_{H_0}\left[\varphi_{\widehat{C},\text{Asymp}}\right] \leq \alpha$. On the other hand, under the alternative hypothesis $H_1 : \mathbb{P}_0 \neq \mathbb{P}_1$, the asymptotic test is consistent as $\lim_{n\to\infty} \mathbb{E}_{H_1}\left[\varphi_{\widehat{C},\text{Asymp}}\right] = 1$.*

Despite its simplicity, the asymptotic approach has no finite sample guarantee on type-1 error control, which motivates an alternative approach based on the permutation principle. In the next subsection we focus on permutation tests and establish the same consistency result.

7.9.2 Permutation tests

In practice, instead of using the asymptotic standard normal null, one often employs permutation tests that can offer exact control of the type-1 error rate. We note that there are two possible ways of applying permutation testing within the classification via sample splitting framework. The methods below differ in the italicized text.

Method 1 (Half-permutation):

- Split data randomly into two halves, call these X^1, Y^1 and X^2, Y^2. Train the classifier on X^1, Y^1, call this f^*. Evaluate accuracy of f^* on X^2, Y^2, call this a^*.

- Repeat P times: *Pool the samples X^2, Y^2 into one bag, randomly permute the samples, and then split it into two parts, X^p, Y^p. Here each part of X^p, Y^p has the same sample size as the corresponding part of X^2, Y^2. Evaluate the accuracy of f^* on this permuted data, call this a^p.*

- Sort all the accuracies $a^*, a^1, ..., a^P$ and denote their order statistics by $a^{(1)} \leq \ldots \leq a^{(P+1)}$; Let $k \stackrel{\text{def}}{=} \lceil (1-\alpha)(1+P) \rceil$. If $a^* > a^{(k)}$, then reject the null.

Method 2 (Full-permutation):

- Split data randomly into two halves, call these X^1, Y^1 and X^2, Y^2. Train the classifier on X^1, Y^1, call this f^*. Evaluate accuracy of f^* on X^2, Y^2, call this a^*.

- Repeat P times: *Pool all samples X^1, Y^1, X^2, Y^2 into one bag, randomly permute the samples, and then split it into 4 parts X^p, Y^p, X'^p, Y'^p. Here each part of X^p, Y^p, X'^p, Y'^p has the same sample size as the corresponding part of X^1, Y^1, X^2, Y^2. Train a new classifier f^p on the first half, evaluate it on the second half, to get accuracy a^p.*

- Sort all the accuracies $a^*, a^1, ..., a^P$ and denote their order statistics by $a^{(1)} \leq \ldots \leq a^{(P+1)}$; Let $k \stackrel{\text{def}}{=} \lceil (1-\alpha)(1+P) \rceil$. If $a^* > a^{(k)}$, then reject the null.

It is worth noting that both methods yield a valid level α test under $H_0 : \mathbb{P}_0 = \mathbb{P}_1$ as a direct consequence of, for example, Theorem 1 in Hemerik and Goeman (2018b). In terms of power, we expect that method 2 is preferred to method 1 as it uses the data more efficiently to determine a cut-off value. In particular permuted accuracies via method 1 can take fewer values (hence the permutation distribution is sparser) than those via method 2, which may result in a more conservative threshold depending on the nominal level. However, we should also note that method 1 has a computational advantage over method 2 since it only requires to re-fit a classifier on the second half of the dataset. Nevertheless the following theorem shows that both methods provide a consistent test under the same assumptions made in Proposition 7.3. Let us denote the permutation test by $\varphi_{\widehat{C},\text{Perm}}$ via either method 1 or method 2 based on classifier $\widehat{C}$. Then our consistency result on $\varphi_{\widehat{C},\text{Perm}}$ is stated as follows.

Theorem 7.6. *Consider the same assumptions made in Proposition 7.3. Then under the null hypothesis $H_0 : \mathbb{P}_0 = \mathbb{P}_1$, we have $\mathbb{E}_{H_0}\left[\varphi_{\widehat{C},\text{Perm}}\right] \leq \alpha$ for each n and d. Under the alternative hypothesis $H_1 : \mathbb{P}_0 \neq \mathbb{P}_1$, the (half or full) permutation test is consistent as $\lim_{n\to\infty} \mathbb{E}_{H_1}\left[\varphi_{\widehat{C},\text{Perm}}\right] = 1$ given that the number of random permutations P is greater than $(1-\alpha)/\alpha$.*

One interesting aspect of the above theorem is that consistency is guaranteed as long as the number of random permutations P is greater than $(1-\alpha)/\alpha$ (e.g., $P \geq 20$ for $\alpha = 0.05$), which is independent of the sample size. We would also like to point out that the permutation test relies on a data-dependent threshold and thus it is more difficult to analyze than the asymptotic test. In Appendix F.3.11, we bound this data-dependent threshold with a more tractable quantity using Markov's inequality with the first two moments of the permuted test statistic. Leveraging this preliminary result, we prove that the permutation critical value cannot exceed the true accuracy in the limit, and this is the critical fact that completes the proof.

7.10 Experiments

In this section, we present several numerical results that support our theoretical analysis. Throughout our simulations (except in Section 7.10.3), we set the sample sizes and the dimension to be $n_0 = n_1 = d = 200$

and compare two multivariate Gaussian or multivariate t-distributions with the same identity covariance matrix. The mean vectors of the two multivariate distributions were chosen to be

$$\mu_0 = (0, \ldots, 0)^\top \quad \text{and} \quad \mu_1 = \frac{\delta}{d^{1/4}} \cdot (1, \ldots, 1)^\top$$

for $\delta \in \{0, 0.05, \ldots, 0.35, 0.40\}$. The simulations were repeated 500 times to estimate the power of each test at significance level $\alpha = 0.05$.

7.10.1 Empirical power vs. theoretical power

In the following experiment, we compare the empirical power of classification accuracy tests with the corresponding theoretical power. For the Gaussian case, we consider the accuracy tests $\varphi_{\Sigma^{-1}}$ and $\varphi_{\widehat{D}^{-1}}$ based on the Fisher's LDA classifier and the naive Bayes classifier, respectively. As specified in the definitions of $\varphi_{\Sigma^{-1}}$ and $\varphi_{\widehat{D}^{-1}}$, the critical values of both tests are based on a normal approximation. Here we split the samples into training and test sets with equal sample sizes so that the power is asymptotically maximized. In this case, the asymptotic power expression for each test is presented in (7.20) and (7.25), respectively. For the case of multivariate t-distributions, we focus on the accuracy test $\varphi_{\Sigma^{-1}}$ and see whether the asymptotic power expression (7.26) approximates its empirical power over different values of degrees of freedom ν.

The results are given in Figure 7.1 and Figure 7.2. From the results, we see that the empirical power almost coincides with the theoretical counterpart especially when δ is not too big (i.e. low SNR regime), which confirms our theoretical analysis. We also see that the accuracy test has higher power when the underlying t-distributions have smaller degrees of freedom, an interesting and initially surprising fact that is again predicted by our theory.

7.10.2 Sample-splitting vs. resubstitution

In the following experiment, we compare the performance of sample-splitting tests with resubstitution accuracy tests under the Gaussian setting. As their name suggests, the resubstitution accuracy tests use resubstitution accuracy estimates as their test statistic. The precise definition of a resubstitution estimate is given in Appendix F.2. We also consider Hotelling's test and its variant proposed by Srivastava and Du (2008) as reference points.The setup is almost the same as the previous experiment except for the choice of critical values. In particular, since the (asymptotic) null distribution of a resubstitution statistic is unknown, the critical values of all tests are determined by the permutation procedure for a fair comparison. Specifically we use method 2, which is described in Section 7.9.2, with 200 random permutations to calibrate critical values.

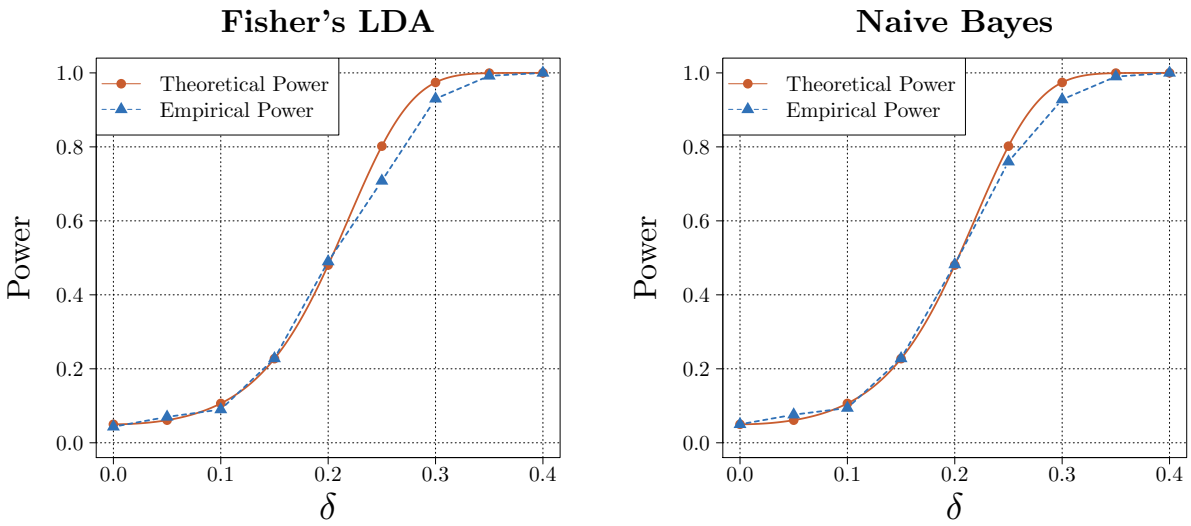

Figure 7.1: Comparisons of the empirical power to our theoretically derived expression for (asymptotic) power under the Gaussian setting. The curves are almost identical especially when the size of δ is not too big, which suggests that our theory under local alternatives accurately predicts power. See Section 7.10.1 for details.

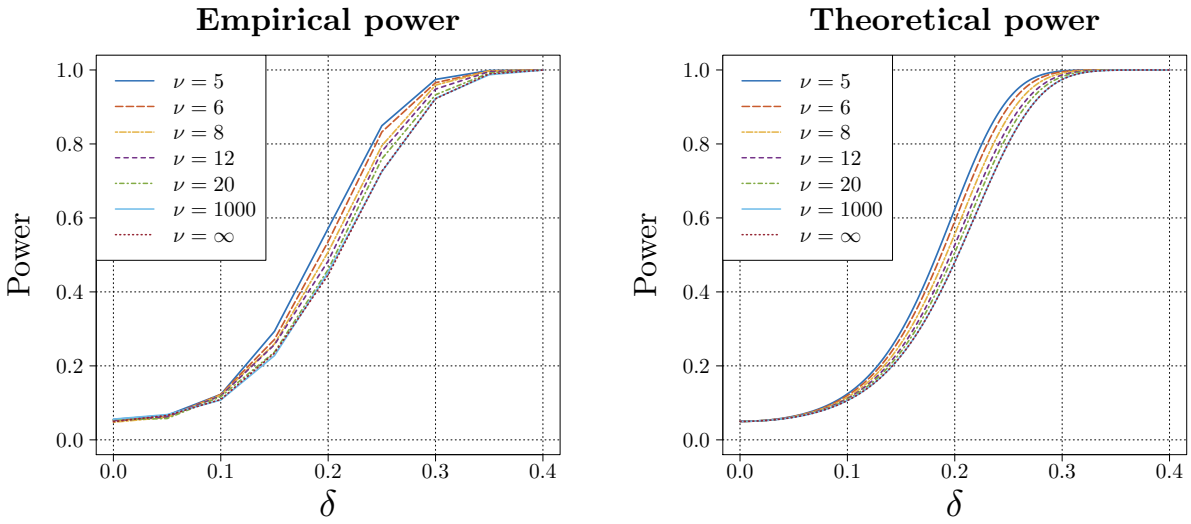

Figure 7.2: The empirical power and theoretical (asymptotic) power of the accuracy test based on Fisher's LDA classifier for comparing multivariate t-distributions with ν degrees of freedom. The empirical power closely follows the corresponding theoretical power over different values of ν. Moreover, predicted by Theorem 7.5, the accuracy test has higher power when the underlying t-distributions have smaller degrees of freedom. See Section 7.10.1 for details.

In the first part, Fisher's LDA is considered as a base line classifier. Then the accuracy is estimated via (i) sample-splitting with $n_{\text{tr}} = n_{\text{te}}$ and (ii) resubstitution. As a reference point, we consider Hotelling's test as it shares the same weight matrix with Fisher's LDA. For both Hotelling's and Fisher's LDA tests, we assume that Σ is known. In the second part, the naive Bayes classifier is considered as a base line classifier

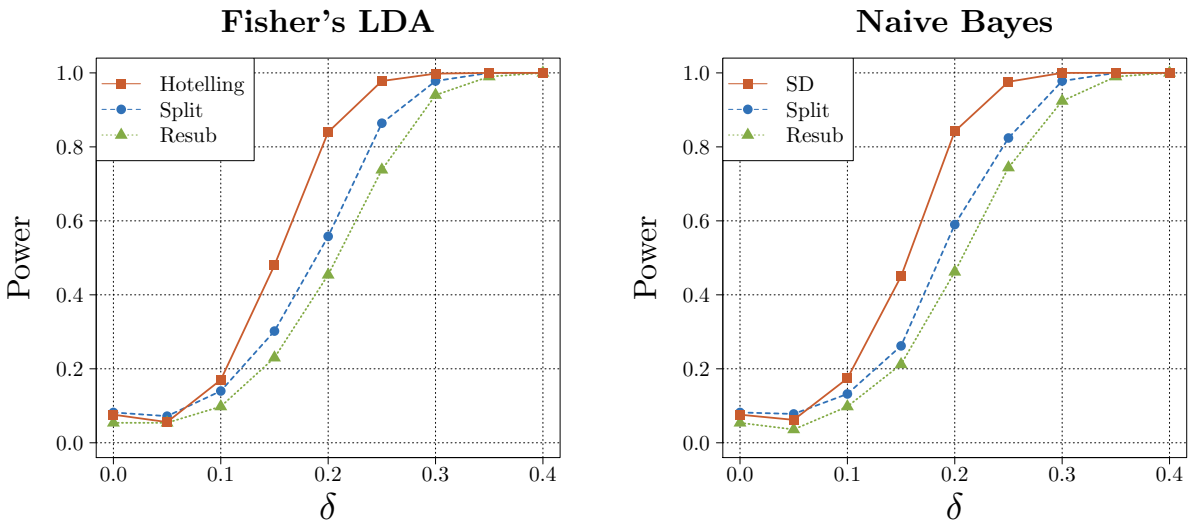

Figure 7.3: Comparisons between sample-splitting (Split) and resubstitution (Resub) tests using Fisher's LDA and naive Bayes classifier. As reference points, we also consider 1) Hotelling's test (Hotelling) and 2) the test based on T_{SD} (SD) in simulations. Under the given scenarios, the sample-splitting tests have higher power than the resubstitution tests but lower power than Hotelling's and SD tests, the latter being predicted by our theory. See Section 7.10.2 for details.

with unknown Σ. We then perform tests based on sample-splitting and resubstitution accuracy statistics defined similarly as before. In this part, we consider T_{SD} given in (7.22) as a reference point since it relies on the inverse of diagonal sample covariance matrix as in the naive Bayes classifier.

From the results presented in Figure 7.3, it stands out that Hotelling's test and its high-dimensional variant are more powerful than the corresponding tests via classification accuracy as we expected. The results also show that the powers of the sample-splitting tests are slightly higher than those of the resubstitution tests in both Fisher's LDA and naive Bayes classifier examples. However additional simulation studies, not presented here, suggest that resubstitution tests tend to be more powerful than sample-splitting tests in low-dimensional settings (or when the sample sizes are relatively small) and thus, at least empirically, neither of them is strictly better than the other under all scenarios. Similar empirical results were observed by Rosenblatt et al. (2016) where they conducted extensive simulation studies to compare the performance of the accuracy tests via resubstitution and 4-fold cross-validation and different versions of Hotelling's test. From their simulation results, one reaches the same conclusion that the accuracy tests tend to have lower power than Hotelling's test against Gaussian mean shift alternatives.

7.10.3 Asymptotic power of Hotelling's Test

In this subsection, we provide numerical support for the asymptotic optimality of Hotelling's test under Gaussian settings with unknown Σ (Theorem 7.1). Here we compare two multivariate Gaussian distributions

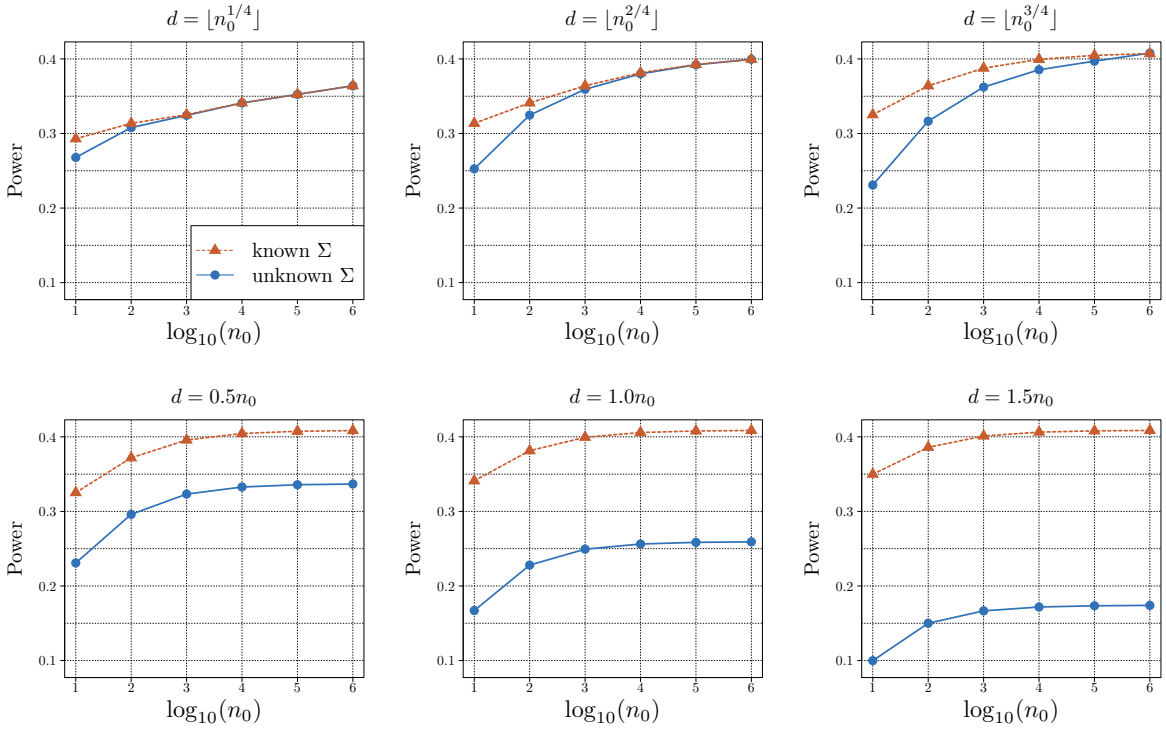

Figure 7.4: Comparisons of the power of the two tests: 1) Hotelling's test φ_H with unknown Σ and 2) Hotelling's test φ_H^* with known Σ at $\alpha = 0.05$ in different asymptotic regimes. These results coincide with our theoretical results in Section 7.4, showing that φ_H has asymptotically the same power as φ_H^* when $d/n \to 0$ (the first row) and it is less powerful when $d/n \to c \in (0, 1)$ (the second row). See Section 7.10.3 for details.

with the mean vectors

$$\mu_0 = \frac{1}{d^{1/4} n_0^{1/2}} \cdot (1, \ldots, 1)^\top \quad \text{and} \quad \mu_1 = -\frac{1}{d^{1/4} n_0^{1/2}} \cdot (1, \ldots, 1)^\top$$

and the identity covariance matrix. In this case, by setting $n_0 = n_1$, the asymptotic minimax power tends to be constant as in (7.8). Now we consider six different asymptotic regimes: i) $d = \lfloor n_0^{1/4} \rfloor$, ii) $d = \lfloor n_0^{2/4} \rfloor$, iii) $d = \lfloor n_0^{3/4} \rfloor$, iv) $d = 0.5n_0$, v) $d = 1.0n_0$ and vi) $d = 1.5n_0$. According to Theorem 7.1, Hotelling's test with unknown Σ (denoted by φ_H) obtains asymptotically the same power as the minimax optimal test (denoted by φ_H^*) in the first three regimes. Whereas, in the last three regimes where d and n are linearly comparable, φ_H becomes less powerful than φ_H^* proved by Bai and Saranadasa (1996). To illustrate this numerically, we increase the sample size by $n_0 \in \{10^1, 10^2, \ldots, 10^6\}$ and compute the power of φ_H^* and φ_H for each n_0. To calculate the power, we use the fact that $\mathbb{E}[1 - \varphi_H^*]$ and $\mathbb{E}[1 - \varphi_H]$ are noncentral χ^2 and F distribution functions evaluated at their critical values, which are $c_{\alpha,d}$ and $q_{\alpha,n,d}$ respectively.

As can be seen in the first row of Figure 7.4, the power of φ_H becomes approximately the same as that of φ_H^* in the first three regimes as n increases. On the other hand, in the last three regimes where

$d/n \to c \in (0, 1)$, we observe significantly different results. Specifically, from the second row of Figure 7.4, it is seen that the power of φ_H is much lower than that of φ_H^* and the gap does not decrease even in large n. This, thereby, supports our argument that φ_H is asymptotically comparable to the minimax optimal test in the case of $d/n \to 0$, but it is underpowered otherwise.

7.11 Conclusions

This chapter provided analyses on the use of classification accuracy as a test statistic for two-sample testing. We started by presenting a fundamental minimax lower bound for high-dimensional two-sample mean testing and showed that Hotelling's test with unknown Σ can be optimal in high-dimensional settings as long as $d/n \to 0$. When $d = O(n)$, we found that two-sample tests via the classification accuracy of various versions of Fisher's LDA (including naive Bayes) have the same power as high-dimensional versions of Hotelling's test in terms of all problem parameters (n, d, δ, Σ), but having worse (but explicit) constants. Beyond linear classifiers, we also proved that both the asymptotic test and the permutation test based on a general classifier are consistent if the limiting value of the true accuracy is higher than chance. This consistency result naturally motivated a more challenging setting in which the Bayes error approaches half while the corresponding accuracy-based test can still have non-trivial power, which is the regime studied in most of this chapter. Under such a challenging regime, it would be interesting to see whether our current results can be extended to non-linear classifiers.

Chapter 8

Minimax optimality of permutation tests

This chapter is adapted from my joint work with Sivaraman Balakrishnan and Larry Wasserman. This work is available on ArXiv (Kim et al., 2020a).

8.1 Introduction

A permutation test is a nonparametric approach to hypothesis testing routinely used in a variety of scientific and engineering applications (e.g. Pesarin and Salmaso, 2010). The permutation test constructs the resampling distribution of a test statistic by permuting the labels of the observations. The resampling distribution, also called the permutation distribution, serves as a reference from which to assess the significance of the observed test statistic. A key property of the permutation test is that it provides exact control of the type I error rate for any test statistic whenever the labels are exchangeable under the null hypothesis (e.g. Hoeffding, 1952). Due to this attractive non-asymptotic property, the permutation test has received considerable attention and has been applied to a wide range of statistical tasks including testing independence, two-sample testing, change point detection, clustering, classification, principal component analysis (see Anderson and Robinson, 2001; Kirch and Steinebach, 2006; Park et al., 2009; Ojala and Garriga, 2010; Zhou et al., 2018).

Once the type I error is controlled, the next concern is the type II error or equivalently power of a test. Despite its increasing popularity and empirical success, the power of the permutation test has yet to be fully understood. A major challenge is to control its random critical value that has an unknown distribution. While some progress has been made as we review in Section 8.1.2, our understanding of the permutation approach is still far from complete, especially in finite-sample scenarios. The purpose of this chapter is to

attempt to fill this gap by developing a general framework for analyzing the non-asymptotic type II error of the permutation test and to demonstrate its efficacy from a minimax point of view.

8.1.1 Alternative approaches and their limitations

We first review a couple of other testing procedures and highlight the advantages of the permutation method. One common approach to determining the critical value of a test is based on the asymptotic null distribution of a test statistic. The validity of a test whose rejection region is calibrated using this asymptotic null distribution is well-studied in the classical regime where the number of parameters is fixed and the sample size goes to infinity. However, it is no longer trivial to justify this asymptotic approach in a complex, high-dimensional setting where numerous parameters can interact in a non-trivial way and strongly influence the behavior of the test statistic. In such a case, the limiting null distribution is perhaps intractable without imposing stringent assumptions. To illustrate the challenge clearly, we consider the two-sample U-statistic U_{n_1,n_2} defined later in Proposition 8.1 for multinomial testing. Here we compute U_{n_1,n_2} based on samples from the multinomial distribution with uniform probabilities. To approximate the null distribution of U_{n_1,n_2}, we perform 1000 Monte Carlo iterations for each bin size $d \in \{5, 100, 10000\}$ while fixing the sample sizes as $n_1 = n_2 = 100$. From the histograms in Figure 8.1, we see that the shape of the null distribution heavily depends on the number of bins d (more generally the probabilities of the multinomial distribution). In particular, the null distribution tends to be more symmetric and sparser as d increases. Since the underlying structure of the distribution is unknown beforehand, Figure 8.1 emphasizes difficulties of approximating the null distribution over different regimes. We also note that the asymptotic approach does not have any finite sample guarantee, which is also true for other data-driven methods including bootstrapping (Efron and Tibshirani, 1994) and subsampling (Politis et al., 1999). In sharp contrast, the permutation approach provides a valid test for any test statistic in any sample size under minimal assumptions. Furthermore, as we shall see, one can achieve minimax power through the permutation test even when a nice limiting null distribution is not available.

Another approach, that is commonly used in theoretical computer science, is based on concentration inequalities (e.g. Chan et al., 2014; Acharya et al., 2014; Bhattacharya and Valiant, 2015; Diakonikolas and Kane, 2016; Canonne et al., 2018). In this approach the threshold of a test is determined using a tail bound of the test statistic under the null hypothesis. Then, owing to the non-asymptotic nature of the concentration bound, the resulting test can control the type I error rate in finite samples. This non-asymptotic approach is more robust to distributional assumptions than the previous asymptotic approach but comes with different challenges. For instance the resulting test tends to be too conservative as it depends on a loose tail bound. More seriously the threshold often relies on unspecified constants and even unknown parameters. By contrast, the permutation approach is entirely data-dependent and tightly controls the type I error rate.

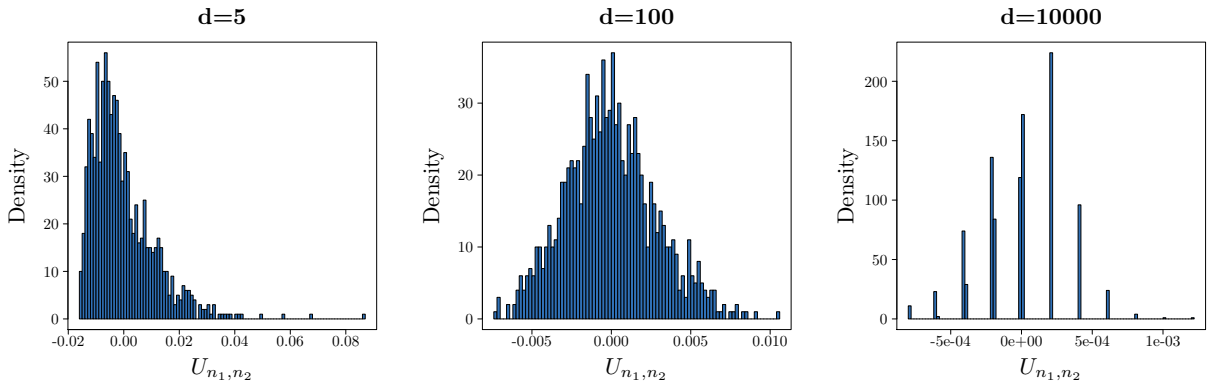

Figure 8.1: Histograms of the U-statistic in Proposition 8.1 calculated under the uniform multinomial null by varying the number of bins d. The plots show that the shape of the null distribution is highly influenced by the bin size and thus illustrate challenges of estimating the null distribution consistently over different scenarios. See Section 8.1.1 for details.

8.1.2 Challenges in power analysis and related work

Having motivated the importance of the permutation approach, we now review the previous studies on permutation tests and also discuss challenges. The large sample power of the permutation test has been investigated by a number of authors including Hoeffding (1952); Robinson (1973); Albers et al. (1976); Bickel and van Zwet (1978). The main result in this line of research indicates that the permutation distribution of a certain test statistic (e.g. Student's t-statistic and F-statistic) approximates its null distribution in large sample scenarios. Moreover this approximation is valid under both the null and local alternatives, which then guarantees that the permutation test is asymptotically as powerful as the test based on the asymptotic null distribution. In addition to these findings, power comparisons between permutation and bootstrap tests have been made by Romano (1989); Janssen and Pauls (2003); Janssen (2005) and among others. We also mention that the robustness property of permutation tests to the exchangeability condition has been investigated by Romano (1990); Chung and Romano (2013); Pauly et al. (2015); Chung and Romano (2016b); DiCiccio and Romano (2017).

However the previous analysis of power, which heavily relies on classical asymptotic theory, is not easily generalized to more complex settings. In particular, it often requires that alternate distributions satisfy certain regularity conditions under which the asymptotic power function is analytically tractable. Due to such restrictions, the focus has been on a limited class of test statistics applied to a relatively small set of distributions. Furthermore, most previous studies have studied the pointwise, instead of uniform, power that holds for any fixed sequence of alternatives but not uniformly over the class of alternatives.

Recently, there has been another line of research studying the power of the permutation test from a non-asymptotic point of view (e.g. Albert, 2015, 2019; Kim et al., 2020b, 2019a). This framework, based

on a concentration bound for a permuted test statistic, allows us to study the power in more general and complex settings than the asymptotic approach at the expense of being less precise (mainly in terms of constant factors). The main challenge in the non-asymptotic analysis, however, is to control the random critical value of the test. The distribution of this random critical value is in general difficult to study due to the non-i.i.d. structure of the permuted test statistic. Several attempts have been made to overcome such difficulty focusing on linear-type statistics (Albert, 2019), regressor-based statistics (Kim et al., 2019a), the Cramér–von Mises statistic (Kim et al., 2020b), maximum-type kernel-based statistics (Kim, 2019) and classification accuracy statistic (Kim et al., 2019b). Our work contributes to this line of research by developing some general tools for studying the finite-sample performance of permutation tests with a specific focus on degenerate U-statistics.

Concurrent with our work, Berrett et al. (2020) also develop results for the permutation test based on a degenerate U-statistic. While focusing on independence testing, Berrett et al. (2020) prove that one cannot hope to have a valid independence test that is uniformly powerful over alternatives in the L_2 distance. The authors then impose Sobolev-type smoothness conditions as well as boundedness conditions on density functions under which the proposed permutation test is minimax rate optimal in the L_2 distance.

Throughout this chapter, we distinguish the L_p distance from the ℓ_p distance — the former is defined with respect to Lebesgue measure and the latter is defined with respect to the counting measure.

8.1.3 Overview of our results

In this chapter we take the non-asymptotic point of view as in Albert (2015) and establish general results to shed light on the power of permutation tests under a variety of scenarios. To concretely demonstrate the efficacy of our results, we focus on two canonical testing problems: 1) *two-sample testing* and 2) *independence testing*, for which the permutation approach rigorously controls the type I error rate (Section 8.2 for specific settings). These topics have been explored by a number of researchers across diverse fields including statistics and computer science and several optimal tests have been proposed in the minimax sense (e.g. Chan et al., 2014; Bhattacharya and Valiant, 2015; Diakonikolas and Kane, 2016; Arias-Castro et al., 2018). Nevertheless the existing optimal tests are mostly of theoretical interest, depending on loose or practically infeasible critical values. Motivated by this gap between theory and practice, the primary goal of this study is to introduce permutation tests that tightly control the type I error rate and have the same optimality guarantee as the existing optimal tests.

We summarize the major contributions of this chapter and contrast them with the previous studies as follows:

- **Two moments method (Lemma 8.0.1).** Leveraging the quantile approach introduced by Fromont et al. (2013) (see Section 8.3 for details), we first present a general sufficient condition under which

the permutation test has non-trivial power. This condition only involves the first two moments of a test statistic, hence called the *two moments method*. To make this general condition more concrete, we consider degenerate U-statistics for two-sample testing and independence testing, respectively, and provide simple moment conditions that ensure that the resulting permutation test has non-trivial power for each testing problem. We then illustrate the efficacy of our results with concrete examples.

- **Multinomial testing (Proposition 8.1 and Proposition 8.4).** One example that we focus on is multinomial testing in the ℓ_2 distance. Chan et al. (2014) study the multinomial two-sample problem in the ℓ_2 distance but with some unnecessary conditions (e.g. equal sample size, Poisson sampling, known squared norms etc). We remove these conditions and propose a permutation test that is minimax rate optimal for the two-sample problem. Similarly we introduce a minimax optimal test for independence testing in the ℓ_2 distance based on the permutation procedure.

- **Density testing (Proposition 8.3 and Proposition 8.7).** Another example that we focus on is density testing for Hölder classes. The two-sample problem for Hölder densities has been studied by Arias-Castro et al. (2018) where the authors propose an optimal test in the minimax sense. However their test depends on a loose critical value and also assumes equal sample sizes. We propose an alternative test based on the permutation procedure without such restrictions and show that it achieves the same minimax optimality. We also contribute to the literature by presenting an optimal permutation test for independence testing over Hölder classes.

- **Combinatorial concentration inequalities (Theorem 8.3, Theorem 8.4 and Theorem 8.5).** Although our two moments method is general, it might be a sub-optimal in terms of the dependence on a nominal level α. Focusing on degenerate U-statistics, we improve the dependence on α from polynomial to logarithmic with some extra assumptions. To do so, we develop combinatorial concentration inequalities inspired by the symmetrization trick (Duembgen, 1998) and Hoeffding's average (Hoeffding, 1963). We apply the developed inequalities to introduce adaptive tests to unknown smoothness parameters at the cost of $\log\log n$ factor. In contrast to the previous studies (e.g. Chatterjee, 2007; Bercu et al., 2015; Albert, 2019) that are restricted to simple linear statistics, the proposed combinatorial inequalities are for degenerate U-statistics, which have potential applications beyond the problems in this chapter (e.g. concentration inequalities under sampling without replacement).

In addition to the testing problems mentioned above, we also contribute to multinomial testing problems in the ℓ_1 distance (e.g. Chan et al., 2014; Bhattacharya and Valiant, 2015; Diakonikolas and Kane, 2016). First we revisit the chi-square test for multinomial two-sample testing considered in Chan et al. (2014) and show that the test based on the same test statistic but calibrated by the permutation procedure is also minimax rate optimal under Poisson sampling (Theorem 8.6). Next, motivated by the flattening idea in

Diakonikolas and Kane (2016), we introduce permutation tests based on weighted U-statistics and prove their minimax rate optimality for multinomial testing in the ℓ_1 distance (Proposition 8.10 and Proposition 8.11). Lastly, building on the recent work of Meynaoui et al. (2019), we analyze the permutation tests based on the maximum mean discrepancy (Gretton et al., 2012) and the Hilbert–Schmidt independence criterion (Gretton et al., 2005) for two-sample and independence testing, respectively, and illustrate their performance over certain function classes.

8.1.4 Outline of the paper

The remainder of the paper is organized as follows. Section 8.2 describes the problem setting and provides some background on the permutation procedure and minimax optimality. In Section 8.3, we give a general condition based on the first two moments of a test statistic under which the permutation test has non-trivial power. We concretely illustrate this condition using degenerate U-statistics for two-sample testing in Section 8.4 and for independence testing in Section 8.5. Section 8.6 is devoted to combinatorial concentration bounds for permuted U-statistics. Building on these results, we propose adaptive tests to unknown smoothness parameters in Section 8.7. The proposed framework is further demonstrated using more sophisticated statistics in Section 8.8. We present some simulation results that justify the permutation approach in Section 8.9 before concluding the paper in Section 8.10. Additional results including concentration bounds for permuted linear statistics and the proofs omitted from the main text are provided in the appendices.

Notation. We use the notation $X \overset{d}{=} Y$ to denote that X and Y have the same distribution. The set of all possible permutations of $\{1, \ldots, n\}$ is denoted by $\mathbf{\Pi}_n$. For two deterministic sequences a_n and b_n, we write $a_n \asymp b_n$ if a_n/b_n is bounded away from zero and ∞ for large n. For integers p, q such that $1 \leq q \leq p$, we let $(p)_q = p(p-1)\cdots(p-q+1)$. We use $\mathbf{i}_q^p$ to denote the set of all q-tuples drawn without replacement from the set $\{1, \ldots, p\}$. $C, C_1, C_2, \ldots$, refer to positive absolute constants whose values may differ in different parts of the paper. We denote a constant that might depend on fixed parameters $\theta_1, \theta_2, \theta_3, \ldots$ by $C(\theta_1, \theta_2, \theta_3, \ldots)$. Given positive integers p and q, we define $\mathbb{S}_p := \{1, \ldots, p\}$ and similarly $\mathbb{S}_{p,q} := \{1, \ldots, p\} \times \{1, \ldots, q\}$.

8.2 Background

We start by formulating the problem of interest. Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two disjoint sets of distributions (or pairs of distributions) on a common measurable space. We are interested in testing whether the underlying data generating distributions belong to $\mathcal{P}_0$ or $\mathcal{P}_1$ based on mutually independent samples $\mathcal{X}_n := \{X_1, \ldots, X_n\}$. Two specific examples of $\mathcal{P}_0$ and $\mathcal{P}_1$ are:

1. **Two-sample testing.** Let (P_Y, P_Z) be a pair of distributions that belongs to a certain family of pairs of distributions $\mathcal{P}$. Suppose we observe $\mathcal{Y}_{n_1} := \{Y_1, \ldots, Y_{n_1}\} \overset{i.i.d.}{\sim} P_Y$ and, independently, $\mathcal{Z}_{n_2} := \{Z_1, \ldots, Z_{n_2}\} \overset{i.i.d.}{\sim} P_Z$ and denote the pooled samples by $\mathcal{X}_n := \mathcal{Y}_{n_1} \cup \mathcal{Z}_{n_2}$. Given the samples, two-sample testing is concerned with distinguishing the hypotheses:

$$H_0 : P_Y = P_Z \quad \text{versus} \quad H_1 : \delta(P_Y, P_Z) \geq \epsilon_{n_1,n_2},$$

where $\delta(P_Y, P_Z)$ is a certain distance between P_Y and P_Z and $\epsilon_{n_1,n_2} > 0$. In this case, $\mathcal{P}_0$ is the set of $(P_Y, P_Z) \in \mathcal{P}$ such that $P_Y = P_Z$, whereas $\mathcal{P}_1 := \mathcal{P}_1(\epsilon_{n_1,n_2})$ is another set of $(P_Y, P_Z) \in \mathcal{P}$ such that $\delta(P_Y, P_Z) \geq \epsilon_{n_1,n_2}$.

2. **Independence testing.** Let P_{YZ} be a joint distribution of Y and Z that belongs to a certain family of distributions $\mathcal{P}$. Let $P_Y P_Z$ denote the product of their marginal distributions. Suppose we observe $\mathcal{X}_n := ((Y_1, Z_1), \ldots, (Y_n, Z_n)) \overset{i.i.d.}{\sim} P_{YZ}$. Given the samples, the hypotheses for testing independence are

$$H_0 : P_{YZ} = P_Y P_Z \quad \text{versus} \quad H_1 : \delta(P_{YZ}, P_Y P_Z) \geq \epsilon_n,$$

where $\delta(P_{YZ}, P_Y P_Z)$ is a certain distance between P_{YZ} and $P_Y P_Z$ and $\epsilon_n > 0$. In this case, $\mathcal{P}_0$ is the set of $P_{YZ} \in \mathcal{P}$ such that $P_{YZ} = P_Y P_Z$, whereas $\mathcal{P}_1 := \mathcal{P}_1(\epsilon_n)$ is another set of $P_{YZ} \in \mathcal{P}$ such that $\delta(P_{YZ}, P_Y P_Z) \geq \epsilon_n$.

Let us consider a generic test statistic $T_n := T_n(\mathcal{X}_n)$, which is designed to distinguish between the null and alternative hypotheses based on $\mathcal{X}_n$. Given a critical value c_n and pre-specified constants $\alpha \in (0, 1)$ and $\beta \in (0, 1 - \alpha)$, the problem of interest is to find sufficient conditions on $\mathcal{P}_0$ and $\mathcal{P}_1$ under which the type I and II errors of the test $\mathbb{1}(T_n > c_n)$ are uniformly bounded as

$$\begin{aligned} &\bullet \text{ Type I error: } \sup_{P \in \mathcal{P}_0} \mathbb{P}_P^{(n)} (T_n > c_n) \leq \alpha, \\ &\bullet \text{ Type II error: } \sup_{P \in \mathcal{P}_1} \mathbb{P}_P^{(n)} (T_n \leq c_n) \leq \beta. \end{aligned} \tag{8.1}$$

Our goal is to control these uniform (rather than pointwise) errors based on data-dependent critical values determined by the permutation procedure.

8.2.1 Permutation procedure

This section briefly overviews the permutation procedure and its well-known theoretical properties, referring readers to Lehmann and Romano (2006); Pesarin and Salmaso (2010) for more details. Let us begin with

some notation. Given a permutation $\pi := (\pi_1, \ldots, \pi_n) \in \mathbf{\Pi}_n$, we denote the permuted version of $\mathcal{X}_n$ by $\mathcal{X}_n^\pi$, that is, $\mathcal{X}_n^\pi := \{X_{\pi_1}, \ldots, X_{\pi_n}\}$. For the case of independence testing, $\mathcal{X}_n^\pi$ is defined by permuting the second variable Z, i.e. $\mathcal{X}_n^\pi := \{(Y_1, Z_{\pi_1}), \ldots, (Y_n, Z_{\pi_n})\}$. We write $T_n^\pi := T_n(\mathcal{X}_n^\pi)$ to denote the test statistic computed based on $\mathcal{X}_n^\pi$. Let $F_{T_n^\pi}(t)$ be the permutation distribution function of T_n^π defined as

$$F_{T_n^\pi}(t) := M_n^{-1} \sum\nolimits_{\pi \in \mathbf{\Pi}_n} \mathbb{1}\{T_n(\mathcal{X}_n^\pi) \leq t\}.$$

Here M_n denotes the cardinality of $\mathbf{\Pi}_n$. We write the $1-\alpha$ quantile of $F_{T_n^\pi}$ by $c_{1-\alpha,n}$ defined as

$$c_{1-\alpha,n} := \inf\{t : F_{T_n^\pi}(t) \geq 1-\alpha\}. \tag{8.2}$$

Given the quantile $c_{1-\alpha,n}$, the permutation test rejects the null hypothesis when $T_n > c_{1-\alpha,n}$. This choice of the critical value provides finite-sample type I error control under the permutation-invariant assumption (or exchangeability). In more detail, the distribution of $\mathcal{X}_n$ is said to be permutation invariant if $\mathcal{X}_n$ and $\mathcal{X}_n^\pi$ have the same distribution whenever the null hypothesis is true. This permutation-invariance is easily met under the settings of the two-sample and independence testing problems. When permutation-invariance holds, it is well-known that the permutation test $\mathbb{1}(T_n > c_{1-\alpha,n})$ is level α and possibly exact by randomizing the test function (see e.g. Hoeffding, 1952; Lehmann and Romano, 2006; Hemerik and Goeman, 2018a).

Remark 8.1 (Computational aspects)**.** Exact calculation of the critical value (8.2) is computationally prohibitive except for small sample sizes. Therefore it is common practice to use Monte-Carlo simulations to approximate the critical value (e.g. Romano and Wolf, 2005). We note that this approximation error can be made arbitrary small by taking a sufficiently large number of Monte-Carlo samples. This argument may be formally justified by using Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956). Hence, while we focus on the exact permutation procedure, all of our results can be extended, in a straightforward manner, to its Monte-Carlo counterpart with a sufficiently large number of Monte-Carlo samples.

8.2.2 Minimax optimality

Another aim of this chapter is to show that the sufficient conditions for the error bounds in (8.1) are indeed necessary in some applications. We approach this problem from a minimax perspective taken by Ingster (1987). Let us define a test ϕ, which is a Borel measurable map, $\phi : \mathcal{X}_n \mapsto \{0, 1\}$. For a class of null distributions $\mathcal{P}_0$, we denote the set of all level α tests by

$$\Phi_{n,\alpha} := \left\{ \phi : \sup_{P \in \mathcal{P}_0} \mathbb{P}_P^{(n)}(\phi = 1) \leq \alpha \right\}.$$

Consider a class of alternative distributions $\mathcal{P}_1(\epsilon_n)$ associated with a positive sequence ϵ_n. Two specific examples of this class of interest are $\mathcal{P}_1(\epsilon_{n_1,n_2}) := \{(P_Y, P_Z) \in \mathcal{P} : \delta(P_Y, P_Z) \geq \epsilon_{n_1,n_2}\}$ for two-sample testing and $\mathcal{P}_1(\epsilon_n) := \{P_{YZ} \in \mathcal{P} : \delta(P_{YZ}, P_Y P_Z) \geq \epsilon_n\}$ for independence testing. Given $\mathcal{P}_1(\epsilon_n)$, the maximum type II error of a test $\phi \in \Phi_{n,\alpha}$ is

$$R_{n,\epsilon_n}(\phi) := \sup_{P \in \mathcal{P}_1(\epsilon_n)} \mathbb{P}_P^{(n)}(\phi = 0),$$

and the minimax risk is defined as

$$R^\dagger_{n,\epsilon_n} := \inf_{\phi \in \Phi_{n,\alpha}} R_n(\phi).$$

The minimax risk is frequently investigated via the minimum separation (or the critical radius), which is the smallest ϵ_n such that type II error becomes non-trivial. Formally, for some fixed $\beta \in (0, 1-\alpha)$, the minimum separation is defined as

$$\epsilon_n^\dagger := \inf\Big\{\epsilon_n : R^\dagger_{n,\epsilon_n} \leq \beta\Big\}.$$

A test $\phi \in \Phi_{n,\alpha}$ is called minimax rate optimal if $R_{n,\epsilon_n}(\phi) \leq \beta$ for some $\epsilon_n \asymp \epsilon_n^\dagger$. With this definition in place, we demonstrate minimax rate optimality of permutation tests in various scenarios.

8.3 A general strategy with first two moments

In this section, we discuss a general strategy for studying the testing errors of a permutation test based on the first two moments of a test statistic. As mentioned earlier, the permutation test is level α as long as permutation-invariance holds under the null hypothesis. Therefore we focus on the type II error rate and provide sufficient conditions under which the error bounds given in (8.1) are fulfilled. The previous approach to the non-asymptotic power analysis, reviewed in Section 8.1.1, hinges on a non-random critical value and thus it does not directly apply to the permutation test. To bridge the gap, we consider a deterministic quantile value that serves as a proxy for the permutation threshold $c_{1-\alpha,n}$. More precisely, let $q_{1-\gamma,n}$ be the $1-\gamma$ quantile of the distribution of the random critical value $c_{1-\alpha,n}$. Then by splitting the cases into $\{c_{1-\alpha,n} \leq q_{1-\gamma,n}\}$ and $\{c_{1-\alpha,n} > q_{1-\gamma,n}\}$ and using the definition of the quantile, it can be shown that the type II error of the permutation test is less than or equal to

$$\sup_{P \in \mathcal{P}_1} \mathbb{P}_P\left(T_n \leq c_{1-\alpha,n}\right) \leq \sup_{P \in \mathcal{P}_1} \mathbb{P}_P\left(T_n \leq q_{1-\gamma,n}\right) + \gamma.$$

Consequently, if one succeeds in showing that $\sup_{P\in\mathcal{P}_1}\mathbb{P}_P\left(T_n \le q_{1-\gamma,n}\right) \le \gamma'$ with γ' such that $\gamma'+\gamma \le \beta$, then the type II error of the permutation test is bounded by β as desired. This quantile approach to dealing with a random threshold is not new and has been considered by Fromont et al. (2013) to study the power of a kernel-based test via a wild bootstrap method. In the next lemma, we build on this quantile approach and study the testing errors of the permutation test based on an arbitrary test statistic. Here and hereafter, we denote the expectation and variance of T_n^π with respect to the permutation distribution by $\mathbb{E}_\pi[T_n^\pi|\mathcal{X}_n]$ and $\mathrm{Var}_\pi[T_n^\pi|\mathcal{X}_n]$, respectively.

Lemma 8.0.1 (Two moments method)**.** *Suppose that for each permutation $\pi \in \mathbf{\Pi}_n$, T_n and T_n^π have the same distribution under the null hypothesis. Given pre-specified error rates $\alpha \in (0,1)$ and $\beta \in (1-\alpha)$, assume that for any $P \in \mathcal{P}_1$,*

$$\begin{aligned}\mathbb{E}_P[T_n] \;\ge\; \mathbb{E}_P[\mathbb{E}_\pi\{T_n^\pi|\mathcal{X}_n\}] &+ \sqrt{\frac{3\mathrm{Var}_P[\mathbb{E}_\pi\{T_n^\pi|\mathcal{X}_n\}]}{\beta}} \\ &+ \sqrt{\frac{3\mathrm{Var}_P[T_n]}{\beta}} + \sqrt{\frac{3\mathbb{E}_P[\mathrm{Var}_\pi\{T_n^\pi|\mathcal{X}_n\}]}{\alpha\beta}}.\end{aligned} \tag{8.3}$$

Then the permutation test $\mathbb{1}(T_n > c_{1-\alpha,n})$ controls the type I and II error rates as in (8.1).

The proof of this general statement follows by simple set algebra along with Markov and Chebyshev's inequalities. The details can be found in Appendix G.4. At a high-level, the sufficient condition (8.3) roughly says that if the expected value of T_n (say, signal) is much larger than the expected value of the permuted statistic T_n^π (say, baseline) as well as the variances of T_n and T_n^π (say, noise), then the permutation test can have non-trivial power greater than the nominal level. We provide an illustration of Lemma 8.0.1 in Figure 8.2. Suppose further that T_n^π is centered at zero under the permutation law, i.e. $\mathbb{E}_\pi[T_n^\pi|\mathcal{X}_n] = 0$. Then a modification of the proof of Lemma 8.0.1 yields a simpler condition with improved constant factors:

$$\mathbb{E}_P[T_n] \;\ge\; \sqrt{\frac{2\mathrm{Var}_P[T_n]}{\beta}} + \sqrt{\frac{2\mathbb{E}_P[\mathrm{Var}_\pi\{T_n^\pi|\mathcal{X}_n\}]}{\alpha\beta}}. \tag{8.4}$$

In the following sections, we demonstrate the two moments method (Lemma 8.0.1) based on degenerate U-statistics for two-sample and independence testing.

8.4 The two moments method for two-sample testing

This section illustrates the two moments method given in Lemma 8.0.1 for two-sample testing. By focusing on a U-statistic, we first present a general condition that ensures that the type I and II error rates of the

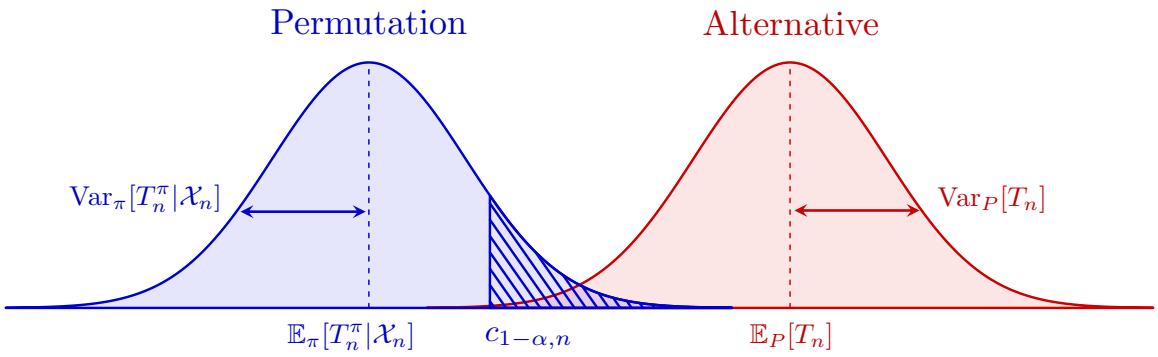

Figure 8.2: An illustration of Lemma 8.0.1. The lemma describes that the major components that determine the power of a permutation test are the mean and the variance of the alternative distribution as well as the permutation distribution. In particular, if the mean of the alternative distribution is sufficiently larger than the other components (on average since the permutation distribution is random), then the permutation test succeeds to reject the null with high probability.

permutation test are uniformly controlled (Theorem 8.1). We then turn to more specific cases of two-sample testing for multinomial distributions and Hölder densities.

Let $g(x, y)$ be a bivariate function, which is symmetric in its arguments, i.e. $g(x, y) = g(y, x)$. Based on this bivariate function, let us define a kernel for a two-sample U-statistic

$$h_{\mathrm{ts}}(y_1, y_2; z_1, z_2) := g(y_1, y_2) + g(z_1, z_2) - g(y_1, z_2) - g(y_2, z_1), \tag{8.5}$$

and write the corresponding U-statistic by

$$U_{n_1,n_2} := \frac{1}{(n_1)_{(2)}(n_2)_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^{n_1}} \sum_{(j_1,j_2)\in \mathbf{i}_2^{n_2}} h_{\mathrm{ts}}(Y_{i_1}, Y_{i_2}; Z_{j_1}, Z_{j_2}). \tag{8.6}$$

Depending on the choice of kernel h_{ts}, the U-statistic includes frequently used two-sample test statistics in the literature such as the maximum mean discrepancy (Gretton et al., 2012) and the energy statistic (Baringhaus and Franz, 2004; Székely and Rizzo, 2004). From the basic properties of U-statistics (e.g. Lee, 1990), it is readily seen that U_{n_1,n_2} is an unbiased estimator of $\mathbb{E}_P[h_{\mathrm{ts}}(Y_1, Y_2; Z_1, Z_2)]$. To describe the main result of this section, let us write the symmetrized kernel by

$$\overline{h}_{\mathrm{ts}}(y_1, y_2; z_1, z_2) := \frac{1}{2!2!} \sum_{(i_1,i_2)\in \mathbf{i}_2^2} \sum_{(j_1,j_2)\in \mathbf{i}_2^2} h_{\mathrm{ts}}(y_{i_1}, y_{i_2}; z_{j_1}, z_{j_2}), \tag{8.7}$$

and define $\psi_{Y,1}(P), \psi_{Z,1}(P)$ and $\psi_{YZ,2}(P)$ by

$$\begin{aligned}
\psi_{Y,1}(P) &:= \mathrm{Var}_P[\mathbb{E}_P\{\overline{h}_{\mathrm{ts}}(Y_1, Y_2; Z_1, Z_2)|Y_1\}], \\
\psi_{Z,1}(P) &:= \mathrm{Var}_P[\mathbb{E}_P\{\overline{h}_{\mathrm{ts}}(Y_1, Y_2; Z_1, Z_2)|Z_1\}], \\
\psi_{YZ,2}(P) &:= \max\{\mathbb{E}_P[g^2(Y_1, Y_2)],\ \mathbb{E}_P[g^2(Y_1, Z_1)],\ \mathbb{E}_P[g^2(Z_1, Z_2)]\}.
\end{aligned} \tag{8.8}$$

The role of $\psi_{Y,1}(P), \psi_{Z,1}(P)$ and $\psi_{YZ,2}(P)$ should be clear in the proof of the following theorem but for now, it is enough to say that these are the key components that upper bound the convergence rate of the variance of U_{n_1,n_2}. By leveraging Lemma 8.0.1, the next theorem presents a sufficient condition that guarantees that the type II error rate of the permutation test based on U_{n_1,n_2} is uniformly bounded by β.

Theorem 8.1 (Two-sample U-statistic)**.** *Suppose that there is a sufficiently large constant $C > 0$ such that*

$$\mathbb{E}_P[U_{n_1,n_2}] \geq C\sqrt{\max\left\{\frac{\psi_{Y,1}(P)}{\beta n_1},\ \frac{\psi_{Z,1}(P)}{\beta n_2},\ \frac{\psi_{YZ,2}(P)}{\alpha\beta}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^2\right\}}, \tag{8.9}$$

for all $P \in \mathcal{P}_1$. Then the type II error of the permutation test over $\mathcal{P}_1$ is uniformly bounded by β, that is

$$\sup_{P\in\mathcal{P}_1} \mathbb{P}_P^{(n_1,n_2)}(U_{n_1,n_2} \leq c_{1-\alpha,n_1,n_2}) \leq \beta.$$

Proof Sketch. Let us give a high-level idea of the proof, while the details are deferred to Appendix G.5. First, by the linearity of expectation, it can be verified that the mean of the permuted U-statistic $U^{\pi}_{n_1,n_2}$ is zero. Therefore it suffices to check condition (8.4). By the well-known variance formula of a two-sample U-statistic (e.g. page 38 of Lee, 1990), we prove in Appendix G.5 that

$$\mathrm{Var}_P[U_{n_1,n_2}] \leq C_1\frac{\psi_{Y,1}(P)}{n_1} + C_2\frac{\psi_{Z,1}(P)}{n_2} + C_3\psi_{YZ,2}(P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^2, \tag{8.10}$$

and this result can be used to bound the first term of condition (8.4). It is worth pointing out that the variance behaves differently under the null and alternative hypotheses. In particular, $\psi_{Y,1}$ and $\psi_{Z,1}$ are zero under the null hypothesis. Hence, in the null case, the third term dominates the variance of U_{n_1,n_2} where we note that $\psi_{YZ,2}$ is a convenient upper bound for the variance of kernel h_{ts}. Intuitively, the permuted U-statistic $U^{\pi}_{n_1,n_2}$ behaves similarly to U_{n_1,n_2} computed based on samples from a certain null distribution (say a mixture of P_Y and P_Z). This implies that the variance of $U^{\pi}_{n_1,n_2}$ is also dominated by the third term in the upper bound (8.10). Having this intuition in mind, we use the symmetric structure of kernel h_{ts} and

prove that

$$\mathbb{E}_P[\mathrm{Var}_\pi\{T_n^\pi|\mathcal{X}_n\}] \le C_4\psi_{YZ,2}(P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^2, \tag{8.11}$$

which is one of our key technical contributions. Based on the previous two bounds in (8.10) and (8.11), we then complete the proof by verifying the sufficient condition (8.4). □

The next two subsections focus on multinomial distributions and Hölder densities and give explicit expressions for condition (8.9). We also demonstrate minimax optimality of permutation tests under the given scenarios.

8.4.1 Two-sample testing for multinomials

Let p_Y and p_Z be multinomial distributions on a discrete domain $\mathbb{S}_d := \{1,\dots,d\}$. Throughout this subsection, we consider the kernel $h_{\mathrm{ts}}(y_1,y_2;z_1,z_2)$ in (8.5) defined with the following bivariate function:

$$g_{\mathrm{Multi}}(x,y) := \sum_{k=1}^{d} \mathbb{1}(x=k)\mathbb{1}(y=k). \tag{8.12}$$

It is straightforward to see that the resulting U-statistic (8.6) is an unbiased estimator of $\|p_Y - p_Z\|_2^2$. Let us denote the maximum between the squared ℓ_2 norms of p_Y and p_Z by

$$b_{(1)} := \max\left\{\|p_Y\|_2^2, \|p_Z\|_2^2\right\}. \tag{8.13}$$

Building on Theorem 8.1, the next result establishes a guarantee on the testing errors of the permutation test under the two-sample multinomial setting.

Proposition 8.1 (Multinomial two-sample testing in ℓ_2 distance)**.** *Let $\mathcal{P}^{(d)}_{\mathrm{Multi}}$ be the set of pairs of multinomial distributions defined on $\mathbb{S}_d$. Let $\mathcal{P}_0 = \{(p_Y,p_Z) \in \mathcal{P}^{(d)}_{\mathrm{Multi}} : p_Y = p_Z\}$ and $\mathcal{P}_1(\epsilon_{n_1,n_2}) = \{(p_Y,p_Z) \in \mathcal{P}^{(d)}_{\mathrm{Multi}} : \|p_Y - p_Z\|_2 \ge \epsilon_{n_1,n_2}\}$ where*

$$\epsilon_{n_1,n_2} \ge C\frac{b_{(1)}^{1/4}}{\alpha^{1/4}\beta^{1/2}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{1/2},$$

for a sufficiently large $C>0$. Consider the two-sample U-statistic U_{n_1,n_2} defined with the bivariate function g_{Multi} given in (8.12). Then the type I and II error rates of the resulting permutation test are uniformly bounded over the classes $\mathcal{P}_0$ and $\mathcal{P}_1$ as in (8.1).

Proof Sketch. We outline the proof of the result, while the details can be found in Appendix G.6. The main technical effort is to show that there exist constants $C_1, C_2, C_3 > 0$ such that

$$\begin{aligned} \psi_{Y,1}(P) &\le C_1 \sqrt{b_{(1)}} \|p_Y - p_Z\|_2^2, \\ \psi_{Z,1}(P) &\le C_2 \sqrt{b_{(1)}} \|p_Y - p_Z\|_2^2, \\ \psi_{YZ,2}(P) &\le C_3 b_{(1)}. \end{aligned} \tag{8.14}$$

These bounds together with Theorem 8.1 imply that if there exists a sufficiently large $C_4 > 0$ such that

$$\|p_Y - p_Z\|_2^2 \ge C_4 \frac{\sqrt{b_{(1)}}}{\alpha^{1/2}\beta} \left(\frac{1}{n_1} + \frac{1}{n_2} \right), \tag{8.15}$$

then the permutation test based on U_{n_1,n_2} has non-trivial power as claimed. □

For the balanced case where $n_1 = n_2$, Chan et al. (2014) prove that no test can have uniform power if ϵ_{n_1,n_2} is of lower order than $b_{(1)}^{1/4} n_1^{-1/2}$. Hence the permutation test in Proposition 8.1 is minimax rate optimal in this balanced setting. The next proposition extends this result to the case of unequal sample sizes and shows that the permutation test is still optimal even for the unbalanced cases.

Proposition 8.2 (Minimum separation for two-sample multinomial testing)**.** *Consider the two-sample testing problem within the class of multinomial distributions* $\mathcal{P}_{\text{Multi}}^{(d)}$ *where the null hypothesis and the alternative hypothesis are* $H_0 : p_Y = p_Z$ *and* $H_1 : \|p_Y - p_Z\|_2 \ge \epsilon_{n_1,n_2}$*. Under this setting and* $n_1 \le n_2$*, the minimum separation satisfies* $\epsilon_{n_1,n_2}^{\dagger} \asymp b_{(1)}^{1/4} n_1^{-1/2}$.

Remark 8.2 (ℓ_1- versus ℓ_2-closeness testing)**.** We note that the minimum separation strongly depends on the choice of metrics. As shown in Bhattacharya and Valiant (2015) and Diakonikolas and Kane (2016), the minimum separation rate for two-sample testing in the ℓ_1 distance is $\max\{d^{1/2} n_2^{-1/4} n_1^{-1/2}, d^{1/4} n_1^{-1/2}\}$ for $n_1 \le n_2$. This rate, in contrast to $b_{(1)}^{1/4} n_1^{-1/2}$, illustrates that the difficulty of ℓ_1-closeness testing depends not only on the smaller sample size n_1 but also on the larger sample size n_2. In Section 8.8.2, we provide a permutation test that is minimax rate optimal in the ℓ_1 distance.

Proof Sketch. We prove Proposition 8.2 indirectly by finding the minimum separation for one-sample multinomial testing. The goal of the one-sample problem is to test whether one set of samples is drawn from a known multinomial distribution. Intuitively the one-sample problem is no harder than the two-sample problem as the former can always be transformed into the latter by drawing another set of samples from the known distribution. This intuition was formalized by Arias-Castro et al. (2018) in which they showed that the minimax risk of the one-sample problem is no larger than that of the two-sample problem (see their

Lemma 1). We prove in Appendix G.7 that the minimum separation for the one-sample problem is of order $b_{(1)}^{1/4} n_1^{-1/2}$ and it thus follows that $b_{(1)}^{1/4} n_1^{-1/2} \lesssim \epsilon^{\dagger}_{n_1,n_2}$. The proof is completed by comparing this lower bound with the upper bound established in Proposition 8.1. □

8.4.2 Two-sample testing for Hölder densities

We next focus on testing for the equality between two density functions under Hölder's regularity condition. Adopting the notation used in Arias-Castro et al. (2018), let $\mathcal{H}_s^d(L)$ be the class of functions $f : [0,1]^d \mapsto \mathbb{R}$ such that

1. $\left|f^{(\lfloor s \rfloor)}(x) - f^{(\lfloor s \rfloor)}(x')\right| \leq L\|x - x'\|^{s-\lfloor s \rfloor}, \quad \forall x, x' \in [0,1]^d,$
2. $\|f^{(s')}\|_\infty \leq L$ for each $s' \in \{1, \ldots, \lfloor s \rfloor\}$,

where $f^{(\lfloor s \rfloor)}$ denotes the $\lfloor s \rfloor$-order derivative of f. Let us write the L_2 norm of $f \in \mathcal{H}_s^d(L)$ by $\|f\|_{L_2}^2 := \int f^2(x)dx$. By letting f_Y and f_Z be the density functions of P_Y and P_Z with respect to Lebesgue measure, we define the set of (P_Y, P_Z), denoted by $\mathcal{P}^{(d,s)}_{\text{Hölder}}$, such that both f_Y and f_Z belong to $\mathcal{H}_s^d(L)$. Under this Hölder density class $\mathcal{P}^{(d,s)}_{\text{Hölder}}$ and $n_1 \leq n_2$, Arias-Castro et al. (2018) establish that for testing $H_0 : f_Y = f_Z$ against $H_1 : \|f_Y - f_Z\|_{L_2} \geq \epsilon_{n_1,n_2}$, the minimum separation rate satisfies

$$\epsilon^{\dagger}_{n_1,n_2} \asymp n_1^{-2s/(4s+d)}. \tag{8.16}$$

We note that this optimal testing rate is faster than the $n^{-s/(2s+d)}$ rate for estimating a Hölder density in the L_2 loss (see Tsybakov, 2009). It is further shown in Arias-Castro et al. (2018) that the optimal rate (8.16) is achieved by the unnormalized chi-square test but with a somewhat loose threshold. Although they recommend a critical value calibrated by permutation in practice, it is unknown whether the resulting test has the same theoretical guarantees. We also note that their testing procedure discards $n_2 - n_1$ observations to balance the sample sizes, which may lead to a less powerful test in practice. Motivated by these limitations, we propose an alternative test for Hölder densities, building on the multinomial permutation test in Proposition 8.1. To implement the multinomial test for continuous data, we first need to discretize the support $[0,1]^d$. We follow the same strategy in Ingster (1987); Arias-Castro et al. (2018); Balakrishnan and Wasserman (2019) and consider bins of equal sizes that partition $[0,1]^d$. In particular, each bin size is set to $\kappa_{(1)}^{-1}$ where $\kappa_{(1)} := \lfloor n_1^{2/(4s+d)} \rfloor$. We then apply the multinomial test in Proposition 8.1 based on the discretized data and have the following theoretical guarantees for density testing.

Proposition 8.3 (Two-sample testing for Hölder densities)**.** *Consider the multinomial test considered in Proposition 8.1 based on the equal-sized binned data described above. For a sufficiently large $C(s,d,L) > 0$,*

consider ϵ_{n_1,n_2} *such that*

$$\epsilon_{n_1,n_2} \geq \frac{C(s,d,L)}{\alpha^{1/4}\beta^{1/2}} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{2s}{4s+d}}.$$

Then for testing $\mathcal{P}_0 = \{(P_Y, P_Z) \in \mathcal{P}^{(d,s)}_{\text{Hölder}} : f_Y = f_Z\}$ *against* $\mathcal{P}_1 = \{(P_Y, P_Z) \in \mathcal{P}^{(d,s)}_{\text{Hölder}} : \|f_Y - f_Z\|_{L_2} \geq \epsilon_{n_1,n_2}\}$*, the type I and II error rates of the resulting permutation test are uniformly controlled as in (8.1).*

The proof of this result uses Proposition 8.1 along with careful analysis of the approximation errors from discretization leveraging Lemma 3 of Arias-Castro et al. (2018). The details can be found in Appendix G.8. We remark that type I error control of the multinomial test follows clearly by the permutation principle, which is not affected by discretization. From the minimum separation rate given in (8.16), it is clear that the proposed test is minimax rate optimal for two-sample testing within Hölder class and it works for both equal and unequal sample sizes without discarding the data. However it is also important to note that the proposed test as well as the test introduced by Arias-Castro et al. (2018) depend on the smoothness parameter s, which is typically unknown. To address this issue, Arias-Castro et al. (2018) build upon the work of Ingster (2000) and propose a Bonferroni-type testing procedure that adapts to this unknown parameter at the cost of $\log n$ factor. In Section 8.7, we improve this logarithmic cost to an iterated logarithmic factor, leveraging combinatorial concentration inequalities developed in Section 8.6.

8.5 The two moments method for independence testing

In this section we present analogous results to those in Section 8.4 for independence testing. We start by introducing a U-statistic for independence testing and establish a general condition under which the permutation test based on the U-statistic controls the type I and II error rates (Theorem 8.2). We then move on to more specific cases of testing for multinomials and Hölder densities in Section 8.5.1 and Section 8.5.2, respectively.

Let us consider two bivariate functions $g_Y(y_1, y_2)$ and $g_Z(z_1, z_2)$, which are symmetric in their arguments. Define a product kernel associated with $g_Y(y_1, y_2)$ and $g_Z(z_1, z_2)$ by

$$\begin{aligned} h_{\text{in}}\{(y_1,z_1),(y_2,z_2),(y_3,z_3),(y_4,z_4)\} &:= \big\{g_Y(y_1,y_2) + g_Y(y_3,y_4) \\ -g_Y(y_1,y_3) - g_Y(y_2,y_4)\big\} &\cdot \big\{g_Z(z_1,z_2) + g_Z(z_3,z_4) - g_Z(z_1,z_3) - g_Z(z_2,z_4)\big\}. \end{aligned} \tag{8.17}$$

For simplicity, we may also write $h_{\text{in}}\{(y_1, z_1), (y_2, z_2), (y_3, z_3), (y_4, z_4)\}$ as $h_{\text{in}}(x_1, x_2, x_3, x_4)$. Given this fourth order kernel, consider a U-statistic defined by

$$U_n := \frac{1}{n_{(4)}} \sum_{(i_1,i_2,i_3,i_4)\in \mathbf{i}_4^n} h_{\text{in}}(X_{i_1}, X_{i_2}, X_{i_3}, X_{i_4}). \tag{8.18}$$

Again, by the unbiasedness property of U-statistics (e.g. Lee, 1990), it is clear that U_n is an unbiased estimator of $\mathbb{E}_P[h_{\text{in}}(X_1, X_2, X_3, X_4)]$. Depending on the choice of kernel h_{in}, the considered U-statistic covers numerous test statistics for independence testing including the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) and distance covariance (Székely et al., 2007). Let $\overline{h}_{\text{in}}(x_1, x_2, x_3, x_4)$ be the symmetrized version of $h_{\text{in}}(x_1, x_2, x_3, x_4)$ given by

$$\overline{h}_{\text{in}}(x_1, x_2, x_3, x_4) := \frac{1}{4!} \sum_{(i_1,i_2,i_3,i_4)\in \mathbf{i}_4^4} h_{\text{in}}(x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4}).$$

In a similar fashion to $\psi_{Y,1}(P), \psi_{Z,1}(P)$ and $\psi_{YZ,2}(P)$, we define $\psi_1'(P)$ and $\psi_2'(P)$ by

$$\begin{aligned}
\psi_1'(P) &:= \text{Var}_P[\mathbb{E}_P\{\overline{h}_{\text{in}}(X_1, X_2, X_3, X_4)|X_1\}], \\
\psi_2'(P) &:= \max\big\{\mathbb{E}_P[g_Y^2(Y_1, Y_2)g_Z^2(Z_1, Z_2)],\ \mathbb{E}_P[g_Y^2(Y_1, Y_2)g_Z^2(Z_1, Z_3)], \\
&\qquad \mathbb{E}_P[g_Y^2(Y_1, Y_2)g_Z^2(Z_3, Z_4)]\big\}.
\end{aligned} \tag{8.19}$$

The following theorem studies the type II error of the permutation test based on U_n.

Theorem 8.2 (U-statistic for independence testing)**.** *Suppose that there is a sufficiently large constant $C > 0$ such that*

$$\mathbb{E}_P[U_n] \geq C\sqrt{\max\left\{\frac{\psi_1'(P)}{\beta n}, \frac{\psi_2'(P)}{\alpha\beta n^2}\right\}},$$

for all $P \in \mathcal{P}_1$. Then the type II error of the permutation test over $\mathcal{P}_1$ is uniformly bounded by β, that is

$$\sup_{P\in\mathcal{P}_1} \mathbb{P}_P^{(n)}(U_n \leq c_{1-\alpha,n}) \leq \beta.$$

Proof Sketch. The proof of Theorem 8.2 proceeds similarly as the proof of Theorem 8.1. Here we present a brief overview of the proof, while the details can be found in Appendix G.9. First of all, the permuted U-statistic U_n^π is centered and it suffices to verify the simplified condition (8.4). To this end, based on the

explicit variance formula of a U-statistic (e.g. page 12 of Lee, 1990), we prove that

$$\mathrm{Var}_P[U_n] \leq C_1 \frac{\psi_1'(P)}{n} + C_2 \frac{\psi_2'(P)}{n^2}. \tag{8.20}$$

Analogous to the case of the two-sample U-statistic, the variance of U_n behaves differently under the null and alternative hypotheses. In particular, under the null hypothesis, $\psi_1'(P)$ becomes zero and thus the second term dominates the upper bound (8.20). Since the permuted U-statistic U_n^π mimics the behavior of U_n under the null, the variance of U_n^π is expected to be similarly bounded. We make this statement precise by proving one of our key technical contributions that

$$\mathbb{E}_P[\mathrm{Var}_\pi\{U_n^\pi | \mathcal{X}_n\}] \leq C_3 \frac{\psi_2'(P)}{n^2}. \tag{8.21}$$

Again, this part of the proof heavily relies on the symmetric structure of kernel h_{in} and the details are deferred to Appendix G.9. Now by combining the established bounds (8.20) and (8.21) together with the sufficient condition (8.4), we can conclude Theorem 8.2. □

In the following subsections, we illustrate Theorem 8.2 in the context of testing multinomial distributions and Hölder densities.

8.5.1 Independence testing for multinomials

We begin with the case of multinomial distributions. Let p_{YZ} denote a multinomial distribution on a product domain $\mathbb{S}_{d_1,d_2} := \{1,\ldots,d_1\} \times \{1,\ldots,d_2\}$ and p_Y and p_Z be its marginal distributions. Let us recall the kernel $h_{\mathrm{in}}(x_1,x_2,x_3,x_4)$ in (8.17) and define it with the following bivariate functions:

$$\begin{aligned} g_{\mathrm{Multi},Y}(y_1,y_2) &:= \sum_{k=1}^{d_1} \mathbb{1}(y_1=k)\mathbb{1}(y_2=k) \quad \text{and} \\ g_{\mathrm{Multi},Z}(z_1,z_2) &:= \sum_{k=1}^{d_2} \mathbb{1}(z_1=k)\mathbb{1}(z_2=k). \end{aligned} \tag{8.22}$$

In this case, the expectation of the U-statistic is $4\|p_{YZ} - p_Y p_Z\|_2^2$. Analogous to the term $b_{(1)}$ in the two-sample case, let us define

$$b_{(2)} := \max\left\{\|p_{YZ}\|_2^2, \|p_Y p_Z\|_2^2\right\}. \tag{8.23}$$

Building on Theorem 8.2, the next result establishes a guarantee on the testing errors of the permutation test for multinomial independence testing.

Proposition 8.4 (Multinomial independence testing in ℓ_2 distance)**.** *Let $\mathcal{P}_{\text{Multi}}^{(d_1,d_2)}$ be the set of multinomial distributions defined on $\mathbb{S}_{d_1,d_2}$. Let $\mathcal{P}_0 = \{p_{YZ} \in \mathcal{P}_{\text{Multi}}^{(d_1,d_2)} : p_{YZ} = p_Y p_Z\}$ and $\mathcal{P}_1(\epsilon_n) = \{p_{YZ} \in \mathcal{P}_{\text{Multi}}^{(d_1,d_2)} : \|p_{YZ} - p_Y p_Z\|_2 \geq \epsilon_n\}$ where*

$$\epsilon_n \geq \frac{C}{\alpha^{1/4}\beta^{1/2}} \frac{b_{(2)}^{1/4}}{n^{1/2}},$$

for a sufficiently large $C > 0$. Consider the U-statistic U_n in (8.18) defined with the bivariate functions $g_{\text{Multi},Y}$ and $g_{\text{Multi},Z}$ given in (8.22). Then, over the classes $\mathcal{P}_0$ and $\mathcal{P}_1$, the type I and II errors of the resulting permutation test are uniformly bounded as in (8.1).

Proof Sketch. We outline the proof of the result, while the details can be found in Appendix G.10. In the proof, we prove that there exist constants $C_1, C_2 > 0$ such that

$$\begin{aligned} \psi_1'(P) &\leq C_1 \sqrt{b_{(2)}} \|p_{YZ} - p_Y p_Z\|_2^2, \\ \psi_2'(P) &\leq C_2 b_{(2)}. \end{aligned} \tag{8.24}$$

These bounds combined with Theorem 8.2 yields that if there exists a sufficiently large $C_3 > 0$ such that

$$\|p_{YZ} - p_Y p_Z\|_2^2 \geq \frac{C_3}{\alpha^{1/2}\beta} \frac{\sqrt{b_{(2)}}}{n}, \tag{8.25}$$

then the type II error of the permutation test can be controlled by β as desired. □

The next proposition asserts that the minimum separation rate for independence testing in the ℓ_2 distance is $\epsilon_n^\dagger \asymp b_{(2)}^{1/4} n^{-1/2}$. This implies that the permutation test based on U_n in Proposition 8.4 is minimax rate optimal in this scenario.

Proposition 8.5 (Minimum separation for multinomial independence testing)**.** *Consider the independence testing problem within the class of multinomial distributions $\mathcal{P}_{\text{Multi}}^{(d_1,d_2)}$ where the null hypothesis and the alternative hypothesis are $H_0 : p_{YZ} = p_Y p_Z$ and $H_1 : \|p_{YZ} - p_Y p_Z\|_2 \geq \epsilon_n$. Under this setting, the minimum separation satisfies $\epsilon_n^\dagger \asymp b_{(2)}^{1/4} n^{-1/2}$.*

The proof of Proposition 8.5 is based on the standard lower bound technique of Ingster (1987) using a uniform mixture of alternative distributions. However, we remark that an extra effort is needed in order to ensure that alternative distributions are proper multinomial distributions. To this end, we carefully perturb the uniform null distribution to generate a mixture of dependent alternative distributions, and use the property of negative association to deal with the dependency. The details can be found in Appendix G.11.

In the next subsection, we turn our attention to the class of Hölder densities and provide similar results of Section 8.4.2 for independence testing.

8.5.2 Independence testing for Hölder densities

Turning to the case of Hölder densities, we leverage the previous multinomial result and establish the minimax rate for independence testing under the Hölder's regularity condition. As in Section 8.4.2, we restrict our attention to functions $f : [0,1]^{d_1+d_2} \mapsto \mathbb{R}$ that satisfy

1. $\left|f^{(\lfloor s \rfloor)}(x) - f^{(\lfloor s \rfloor)}(x')\right| \leq L\|x - x'\|^{s-\lfloor s \rfloor}, \quad \forall x, x' \in [0,1]^{d_1+d_2}$,
2. $\|f^{(s')}\|_\infty \leq L$ for each $s' \in \{1, \ldots, \lfloor s \rfloor\}$.

Let us write $\mathcal{H}_s^{d_1+d_2}(L)$ to denote the class of such functions. We further introduce the class of joint distributions, denoted by $\mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)}$, defined as follows. Let f_{YZ} and $f_Y f_Z$ be the densities of P_{YZ} and $P_Y P_Z$ with respect to Lebesgue measure. Then $\mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)}$ is defined as the set of joint distributions P_{YZ} such that both the joint density and the product density, f_{YZ} and $f_Y f_Z$, belong to $\mathcal{H}_s^{d_1+d_2}(L)$. Consider partitions of $[0,1]^{d_1+d_2}$ into bins of equal size and set the bin size to be $\kappa_{(2)}^{-1}$ where $\kappa_{(2)} = \lfloor n^{2/(4s+d_1+d_2)} \rfloor$. Based on these equal-sized partitions, one may apply the multinomial test for independence provided in Proposition 8.4. Despite discretization, the resulting test has valid level α due to the permutation principle and has the following theoretical guarantees for density testing over $\mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)}$.

Proposition 8.6 (Independence testing for Hölder densities)**.** *Consider the multinomial independence test considered in Proposition 8.4 based on the binned data described above. For a sufficiently large $C(s, d_1, d_2, L) > 0$, consider ϵ_n defined by*

$$\epsilon_n \geq \frac{C(s, d_1, d_2, L)}{\alpha^{1/4}\beta^{1/2}} \left(\frac{1}{n}\right)^{\frac{2s}{4s+d_1+d_2}}.$$

Then for testing $\mathcal{P}_0 = \{P_{YZ} \in \mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)} : f_{YZ} = f_Y f_Z\}$ against $\mathcal{P}_1 = \{P_{YZ} \in \mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)} : \|f_{YZ} - f_Y f_Z\|_{L_2} \geq \epsilon_n\}$, the type I and II errors of the resulting permutation test are uniformly controlled as in (8.1).

The proof of the above result follows similarly to the proof of Proposition 8.3 and can be found in Appendix G.12. Indeed, as shown in the next proposition, the proposed binning-based independence test is minimax rate optimal for the Hölder class density functions. That is, no test can have uniform power when the separation rate ϵ_n is of order smaller than $n^{-4s/(4s+d_1+d_2)}$.

Proposition 8.7 (Minimum separation for independence testing in Hölder class)**.** *Consider the independence testing problem within the class $\mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)}$ in which the null hypothesis and the alternative hypothesis are*

$H_0 : f_{YZ} = f_Y f_Z$ *and* $H_1 : \|f_{YZ} - f_Y f_Z\|_{L_2} \geq \epsilon_n$. *Under this setting, the minimum separation satisfies* $\epsilon_n^\dagger \asymp n^{-2s/(4s+d_1+d_2)}$.

The proof of Proposition 8.7 is again based on the standard lower bound technique by Ingster (1987) and deferred to Appendix G.13. We note that the independence test in Proposition 8.6 hinges on the assumption that the smoothness parameter s is known. To avoid this assumption, we introduce an adaptive test to this smoothness parameter at the cost of $\log\log n$ factor in Section 8.7. A building block for this adaptive result is combinatorial concentration inequalities developed in the next section.

8.6 Combinatorial concentration inequalities

Although the two moments method is broadly applicable, it may not yield sharp results when an extremely small significance level α is of interest (say, α shrinks to zero as n increases). In particular, the sufficient condition (8.3) given by the two moments method has a polynomial dependency on α. In this section, we develop exponential concentration inequalities for permuted U-statistics that allow us to improve this polynomial dependency. To this end, we introduce a novel strategy to couple a permuted U-statistic with i.i.d. Bernoulli or Rademacher random variables, inspired by the symmetrization trick (Duembgen, 1998) and Hoeffding's average (Hoeffding, 1963).

Coupling with i.i.d. random variables. The core idea of our approach is fairly general and based on the following simple observation. Given a random permutation π uniformly distributed over $\mathbf{\Pi}_n$, we randomly switch the order within (π_{2i-1}, π_{2i}) for $i = 1, 2, \ldots, \lfloor n/2 \rfloor$. We denote the resulting permutation by π'. It is clear that π and π' are dependent but identically distributed. The point of introducing this extra permutation π' is that we are now able to associate π' with i.i.d. Bernoulli random variables without changing the distribution. To be more specific, let $\delta_1, \ldots, \delta_{\lfloor n/2 \rfloor}$ be i.i.d. Bernoulli random variables with success probability $1/2$. Then (π'_{2i-1}, π'_{2i}) can be written as

$$(\pi'_{2i-1}, \pi'_{2i}) = \big(\delta_i \pi_{2i-1} + (1-\delta_i)\pi_{2i}, (1-\delta_i)\pi_{2i-1} + \delta_i \pi_{2i}\big) \quad \text{for } i = 1, 2, \ldots, \lfloor n/2 \rfloor.$$

Given that it is easier to work with i.i.d. samples than permutations, the alternative representation of π' gives a nice way to investigate a general permuted statistic. The next subsections provide concrete demonstrations of this coupling approach based on degenerate U-statistics.

8.6.1 Degenerate two-sample U-statistics

We start with the two-sample U-statistic in (8.6). Our strategy is outlined as follows. First, motivated by Hoeffding's average (Hoeffding, 1963), we express the permuted U-statistic as the average of more tractable

statistics. We then link these tractable statistics to quadratic forms of i.i.d. Rademacher random variables based on the coupling idea described before. Finally we apply existing concentration bounds for quadratic forms of i.i.d. random variables to obtain the result in Theorem 8.3.

Let us denote the permuted U-statistic associated with $\pi \in \mathbf{\Pi}_n$ by

$$U^{\pi}_{n_1,n_2} := \frac{1}{(n_1)_{(2)}(n_2)_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^{n_1}} \sum_{(j_1,j_2)\in \mathbf{i}_2^{n_2}} h_{\mathrm{ts}}(X_{\pi_{i_1}}, X_{\pi_{i_2}}; X_{\pi_{n_1+j_1}}, X_{\pi_{n_1+j_2}}). \tag{8.26}$$

By assuming $n_1 \leq n_2$, let $L := \{\ell_1, \ldots, \ell_{n_1}\}$ be a n_1-tuple uniformly drawn without replacement from $\{1, \ldots, n_2\}$. Given L, we introduce another test statistic

$$\widetilde{U}^{\pi,L}_{n_1,n_2} := \frac{1}{(n_1)_{(2)}} \sum_{(k_1,k_2)\in \mathbf{i}_2^{n_1}} h_{\mathrm{ts}}(X_{\pi_{k_1}}, X_{\pi_{k_2}}; X_{\pi_{n_1+\ell_{k_1}}}, X_{\pi_{n_1+\ell_{k_2}}}).$$

By treating L as a random quantity, $U^{\pi}_{n_1,n_2}$ can be viewed as the expected value of $\widetilde{U}^{\pi,L}_{n_1,n_2}$ with respect to L (conditional on other random variables), that is,

$$U^{\pi}_{n_1,n_2} = \mathbb{E}_L[\widetilde{U}^{\pi,L}_{n_1,n_2} | \mathcal{X}_n, \pi]. \tag{8.27}$$

The idea of expressing a U-statistic as the average of more tractable statistics dates back to Hoeffding (1963). The reason for introducing $\widetilde{U}^{\pi,L}_{n_1,n_2}$ is to connect $U^{\pi}_{n_1,n_2}$ with a Rademacher chaos. Recall that $\pi = (\pi_1, \ldots, \pi_n)$ is uniformly distributed over all possible permutations of $\{1, \ldots, n\}$. Therefore, as explained earlier, the distribution of $\widetilde{U}^{\pi,L}_{n_1,n_2}$ does not change even if we randomly switch the order between X_{π_k} and $X_{\pi_{n_1+\ell_k}}$ for $k \in \{1, \ldots, n_1\}$. More formally, recall that $\delta_1, \ldots, \delta_{n_1}$ are i.i.d. Bernoulli random variables with success probability $1/2$. For $k = 1, \ldots, n_1$, define

$$\widetilde{X}_{\pi_k} := \delta_k X_{\pi_k} + (1-\delta_k) X_{\pi_{n_1+\ell_k}} \quad \text{and} \quad \widetilde{X}_{\pi_{n_1+\ell_k}} := (1-\delta_k) X_{\pi_k} + \delta_k X_{\pi_{n_1+\ell_k}}. \tag{8.28}$$

Then it can be seen that $\widetilde{U}^{\pi,L}_{n_1,n_2}$ is equal in distribution to

$$\widetilde{U}^{\pi,L,\delta}_{n_1,n_2} := \frac{1}{(n_1)_{(2)}} \sum_{(k_1,k_2)\in \mathbf{i}_2^{n_1}} h_{\mathrm{ts}}(\widetilde{X}_{\pi_{k_1}}, \widetilde{X}_{\pi_{k_2}}; \widetilde{X}_{\pi_{n_1+\ell_{k_1}}}, \widetilde{X}_{\pi_{n_1+\ell_{k_2}}}).$$

In other words, we link $\widetilde{U}^{\pi,L}_{n_1,n_2}$ to i.i.d. Bernoulli random variables, which are easier to work with. Furthermore, by the symmetry of $g(x,y)$ in its arguments and letting $\zeta_1, \ldots, \zeta_n$ be i.i.d. Rademacher random

variables, one can observe that $\widetilde{U}^{\pi,L,\delta}_{n_1,n_2}$ is equal in distribution to the following Rademacher chaos:

$$\widetilde{U}^{\pi,L,\zeta}_{n_1,n_2} := \frac{1}{(n_1)_{(2)}} \sum_{(k_1,k_2)\in \mathbf{i}_2^{n_1}} \zeta_{k_1}\zeta_{k_2} h_{\mathrm{ts}}(X_{\pi_{k_1}}, X_{\pi_{k_2}}; X_{\pi_{n_1+\ell_{k_1}}}, X_{\pi_{n_1+\ell_{k_2}}}).$$

Consequently, we observe that $\widetilde{U}^{\pi,L}_{n_1,n_2}$ and $\widetilde{U}^{\pi,L,\zeta}_{n_1,n_2}$ are equal in distribution, i.e.

$$\widetilde{U}^{\pi,L}_{n_1,n_2} \overset{d}{=} \widetilde{U}^{\pi,L,\zeta}_{n_1,n_2}. \tag{8.29}$$

We now have all the ingredients ready for obtaining an exponential bound for $U^{\pi}_{n_1,n_2}$. By the Chernoff bound (e.g. Boucheron et al., 2013), for any $\lambda > 0$,

$$\begin{aligned} \mathbb{P}_\pi\left(U^{\pi}_{n_1,n_2} > t|\mathcal{X}_n\right) &\leq e^{-\lambda t}\mathbb{E}_\pi\left[\exp\left(\lambda U^{\pi}_{n_1,n_2}\right)|\mathcal{X}_n\right] \\ &\overset{(i)}{\leq} e^{-\lambda t}\mathbb{E}_{\pi,L}\left[\exp\left(\lambda \widetilde{U}^{\pi,L}_{n_1,n_2}\right)|\mathcal{X}_n\right] \\ &\overset{(ii)}{=} e^{-\lambda t}\mathbb{E}_{\pi,L,\zeta}\left[\exp\left(\lambda \widetilde{U}^{\pi,L,\zeta}_{n_1,n_2}\right)|\mathcal{X}_n\right] \end{aligned} \tag{8.30}$$

where step (i) uses Jensen's inequality together with (8.27) and step (ii) holds from (8.29). Finally, conditional on π and L, we can associate the last equation with the moment generating function of a quadratic form of i.i.d. Rademacher random variables. This quadratic form has been well-studied in the literature through a decoupling argument (e.g. Chapter 6 of Vershynin, 2018), which leads to the following theorem. The remainder of the proof of Theorem 8.3 can be found in Section G.14.

Theorem 8.3 (Concentration of $U^{\pi}_{n_1,n_2}$)**.** *Consider the permuted two-sample U-statistic $U^{\pi}_{n_1,n_2}$ (8.26) and define*

$$\Sigma^2_{n_1,n_2} := \frac{1}{n_1^2(n_1-1)^2} \sup_{\pi\in\mathbf{\Pi}_n} \left\{ \sum_{(i_1,i_2)\in\mathbf{i}_2^{n_1}} g^2(X_{\pi_{i_1}}, X_{\pi_{i_2}}) \right\}.$$

Then, for every $t > 0$ and some constant $C > 0$, we have

$$\mathbb{P}_\pi\left(U^{\pi}_{n_1,n_2} \geq t \ |\mathcal{X}_n\right) \leq \exp\left\{ -C\min\left(\frac{t^2}{\Sigma^2_{n_1,n_2}}, \frac{t}{\Sigma_{n_1,n_2}}\right)\right\}.$$

In our application, it is convenient to have an upper bound for Σ_{n_1,n_2} without involving the supremum operator. One trivial bound, suitable for our purpose, is given by

$$\Sigma^2_{n_1,n_2} \leq \frac{1}{n_1^2(n_1-1)^2} \sum_{(i_1,i_2)\in\mathbf{i}_2^{n}} g^2(X_{i_1}, X_{i_2}). \tag{8.31}$$

The next subsection presents an analogous result for degenerate U-statistics in the context of independence testing.

8.6.2 Degenerate U-statistics for independence testing

Let us recall the U-statistic for independence testing in (8.18) and denote the permuted version by

$$U_n^{\pi} := \frac{1}{n_{(4)}} \sum_{(i_1,i_2,i_3,i_4)\in \mathbf{i}_4^n} h_{\text{in}}\big\{(Y_{i_1}, Z_{\pi_{i_1}}), (Y_{i_2}, Z_{\pi_{i_2}}), (Y_{i_3}, Z_{\pi_{i_3}}), (Y_{i_4}, Z_{\pi_{i_4}})\big\}. \tag{8.32}$$

We follow a similar strategy taken in the previous subsection to obtain an exponential bound for U_n^{π}. To this end, we first introduce some notation. Let $L := \{\ell_1, \ldots, \ell_{\lfloor n/2 \rfloor}\}$ be a $\lfloor n/2 \rfloor$-tuple uniformly sampled without replacement from $\{1, \ldots, n\}$ and similarly $L' := \{\ell'_1, \ldots, \ell'_{\lfloor n/2 \rfloor}\}$ be another $\lfloor n/2 \rfloor$-tuple uniformly sampled without replacement from $\{1, \ldots, n\} \setminus L$. By construction, L and L' are disjoint. Given L and L', we define another test statistic $\widetilde{U}_n^{\pi,L,L'}$ as

$$\widetilde{U}_n^{\pi,L,L'} := \frac{1}{\lfloor n/2 \rfloor_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^{\lfloor n/2 \rfloor}} h_{\text{in}}\big\{\big(Y_{\ell_{i_1}}, Z_{\pi_{\ell_{i_1}}}\big), \big(Y_{\ell_{i_2}}, Z_{\pi_{\ell_{i_2}}}\big), \big(Y_{\ell'_{i_2}}, Z_{\pi_{\ell'_{i_2}}}\big), \big(Y_{\ell'_{i_1}}, Z_{\pi_{\ell'_{i_1}}}\big)\big\}.$$

By treating L and L' as random quantities, U_n^{π} can be viewed as the expected value of $\widetilde{U}_n^{\pi,L,L'}$ with respect to L and L', i.e.

$$U_n^{\pi} = \mathbb{E}_{L,L'}[\widetilde{U}_n^{\pi,L,L'} | \mathcal{X}_n, \pi]. \tag{8.33}$$

From the same reasoning as before, the distribution of $\widetilde{U}_n^{\pi,L,L'}$ does not change even if we randomly switch the order between $Z_{\pi_{\ell_k}}$ and $Z_{\pi_{\ell'_k}}$ for $k = 1, \ldots, \lfloor n/2 \rfloor$, which allows us to introduce i.i.d. Bernoulli random variables with success probability $1/2$. By the symmetry of $g_Y(y_1, y_2)$ and $g_Z(z_1, z_2)$, we may further observe that $\widetilde{U}_n^{\pi,L,L'}$ is equal in distribution to

$$\begin{aligned} \widetilde{U}_n^{\pi,L,L',\zeta} := \; & \frac{1}{\lfloor n/2 \rfloor_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^{\lfloor n/2 \rfloor}} \zeta_{i_1}\zeta_{i_2} \times \\ & h_{\text{in}}\big\{\big(Y_{\ell_{i_1}}, Z_{\pi_{\ell_{i_1}}}\big), \big(Y_{\ell_{i_2}}, Z_{\pi_{\ell_{i_2}}}\big), \big(Y_{\ell'_{i_2}}, Z_{\pi_{\ell'_{i_2}}}\big), \big(Y_{\ell'_{i_1}}, Z_{\pi_{\ell'_{i_1}}}\big)\big\}. \end{aligned} \tag{8.34}$$

Thus, based on the alternative expression of U_n^{π} in (8.33) along with the relationship $\widetilde{U}_n^{\pi,L,L'} \overset{d}{=} \widetilde{U}_n^{\pi,L,L',\zeta}$, we can establish a similar exponential tail bound as in Theorem 8.3 for U_n^{π} as follows.

Theorem 8.4 (Concentration I of U_n^π)**.** *Consider the permuted U-statistic U_n^π (8.32) and define*

$$\Sigma_n^2 := \frac{1}{n^2(n-1)^2} \sup_{\pi \in \mathbf{\Pi}_n} \left\{ \sum_{(i_1,i_2)\in \mathbf{i}_2^n} g_Y^2(Y_{i_1}, Y_{i_2}) g_Z^2(Z_{\pi_{i_1}}, Z_{\pi_{i_2}}) \right\}. \tag{8.35}$$

Then, for every $t > 0$ and some constant $C > 0$, we have

$$\mathbb{P}_\pi \left(U_n^\pi \geq t \mid \mathcal{X}_n \right) \leq \exp\left\{ - C \min\left(\frac{t^2}{\Sigma_n^2}, \frac{t}{\Sigma_n} \right) \right\}.$$

We omit the proof of the result as it follows exactly the same line of the proof of Theorem 8.3. Similar to the upper bound (8.31), Hölder's inequality yields two convenient bounds for Σ_n^2 as

$$\begin{aligned} \Sigma_n^2 &\leq \frac{1}{n^2(n-1)^2} \|\!| g_Z^2 |\!\|_\infty \sum_{(i_1,i_2)\in \mathbf{i}_2^n} g_Y^2(Y_{i_1}, Y_{i_2}) \quad \text{and} \\ \Sigma_n^2 &\leq \frac{1}{n^2(n-1)^2} \sqrt{\sum_{(i_1,i_2)\in \mathbf{i}_2^n} g_Y^4(Y_{i_1}, Y_{i_2})} \sqrt{\sum_{(i_1,i_2)\in \mathbf{i}_2^n} g_Z^4(Z_{i_1}, Z_{i_2})}. \end{aligned} \tag{8.36}$$

At the end of this subsection, we provide an application of Theorem 8.4 to a dependent Rademacher chaos.

A refined version. Although Theorem 8.4 presents a fairly strong exponential concentration of U_n^π, it may lead to a sub-optimal result for independence testing. Indeed, for the minimax result, we want to obtain a similar bound but by replacing the supremum with the average over $\pi \in \mathbf{\Pi}_n$ in (8.35). To this end, we borrow decoupling ideas from Duembgen (1998) and De la Pena and Giné (1999) and present a refined concentration inequality in Theorem 8.5. The proposed bound (8.38) can be viewed as Bernstein-type inequality in a sense that it contains the variance term Λ_n (not depending on the supremum) and maximum term M_n defined as

$$\begin{aligned} \Lambda_n^2 &:= \frac{1}{n^4} \sum_{1\leq i_1,i_2\leq n} \sum_{1\leq j_1,j_2\leq n} g_Y^2(Y_{i_1}, Y_{i_2}) g_Z^2(Z_{j_1}, Z_{j_2}) \quad \text{and} \\ M_n &:= \max_{1\leq i_1,i_2,j_1,j_2\leq n} |g_Y(Y_{i_1}, Y_{i_2}) g_Z(Z_{j_1}, Z_{j_2})|. \end{aligned} \tag{8.37}$$

In particular, the revised inequality would be sharper than the one in Theorem 8.4 especially when Λ_n is much smaller than $n\Sigma_n$.

Theorem 8.5 (Concentration II of U_n^π)**.** *Consider the permuted U-statistic U_n^π (8.32) and recall Λ_n and M_n from (8.37). Then, for every $t > 0$ and some constant $C_1, C_2 > 0$, we have*

$$\mathbb{P}_\pi \left(U_n^\pi \geq t \mid \mathcal{X}_n \right) \leq C_1 \exp\left\{ - C_2 \min\left(\frac{nt}{\Lambda_n}, \frac{nt^{2/3}}{M_n^{3/2}} \right) \right\}. \tag{8.38}$$

Proof Sketch. Here we sketch the proof of the result while the details are deferred to Appendix G.16. Let $\psi(\cdot)$ be a nondecreasing convex function on $[0,\infty)$ and $\Psi(x) = \psi(|x|)$. Based on the equality in (8.33), Jensen's inequality yields

$$\mathbb{E}_\pi[\Psi(\lambda U_n^\pi)|\mathcal{X}_n] \leq \mathbb{E}_{\pi,L,L',\zeta}\big[\Psi\big(\lambda \widetilde{U}_n^{\pi,L,L',\zeta}\big)|\mathcal{X}_n\big],$$

where $\widetilde{U}_n^{\pi,L,L',\zeta}$ can be recalled from (8.34). Let π' be i.i.d. copy of permutation π. Then, by letting $m = \lfloor n/2 \rfloor$ and observing that (i) $\{\zeta_i\}_{i=1}^m \overset{d}{=} \{\zeta_i\zeta_{i+m}\}_{i=1}^m$ and (ii) $\{L, L'\} \overset{d}{=} \{\pi'_1, \ldots, \pi'_{2m}\}$, we have

$$\begin{aligned}\widetilde{U}_n^{\pi,L,L',\zeta} \overset{d}{=} \widetilde{U}_n^{\pi,\pi',\zeta} := \frac{1}{m_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^m} \zeta_{i_1}\zeta_{i_2}\zeta_{i_1+m}\zeta_{i_2+m} \times \\ h_{\text{in}}\{(Y_{\pi'_{i_1}}, Z_{\pi_{i_1}}), (Y_{\pi'_{i_2}}, Z_{\pi_{i_2}}), (Y_{\pi'_{i_2+m}}, Z_{\pi_{\pi_{i_2+m}}}), (Y_{\pi'_{i_1+m}}, Z_{\pi_{i_1+m}})\}.\end{aligned}$$

Next denote the decoupled version of π by $\widetilde{\pi} := (\widetilde{\pi}_1, \ldots, \widetilde{\pi}_n)$ whose components are independent and identically distributed as π_1. Let $\widetilde{\pi}'$ be i.i.d. copy of $\widetilde{\pi}$. Building on the decoupling idea of Duembgen (1998), our proof proceeds by replacing π, π' in $\widetilde{U}_n^{\pi,\pi',\zeta}$ with $\widetilde{\pi}, \widetilde{\pi}'$. If this decoupling step succeeds, then we can view the corresponding U-statistic as a second order degenerate U-statistic of i.i.d. random variables (conditional on $\mathcal{X}_n$). We are then able to apply concentration inequalities for degenerate U-statistics in De la Pena and Giné (1999) to finish the proof. □

Dependent Rademacher chaos. To illustrate the efficacy of Theorem 8.4, let us consider a Rademacher chaos under sampling without replacement, which has been recently studied by Hodara and Reynaud-Bouret (2019). To describe the problem, let $\widetilde{\zeta}_1, \ldots, \widetilde{\zeta}_n$ be dependent Rademacher random variables such that $\sum_{i=1}^n \widetilde{\zeta}_i = 0$ where n is assumed to be even. For real numbers $\{a_{i,j}\}_{i,j=1}^n$, the Rademacher chaos under sampling without replacement is given by

$$T_{\text{Rad}} := \sum_{(i_1,i_2)\in \mathbf{i}_2^n} \widetilde{\zeta}_{i_1}\widetilde{\zeta}_{i_2} a_{i_1,i_2}.$$

Hodara and Reynaud-Bouret (2019) present two exponential concentration inequalities for T_{Rad} based on the coupling argument introduced by Chung and Romano (2013). Intuitively, T_{Rad} should behave like i.i.d. Rademacher chaos, replacing $\{\widetilde{\zeta}_i\}_{i=1}^n$ with $\{\zeta_i\}_{i=1}^n$, at least in the large sample size. Both of their results, however, do not fully recover a well-known concentration bound for i.i.d. Rademacher chaos (e.g. Corollary 3.2.6 of De la Pena and Giné, 1999); namely,

$$\mathbb{P}\Big\{\Big|\sum\nolimits_{(i_1,i_2)\in \mathbf{i}_2^n} \zeta_{i_1}\zeta_{i_2} a_{i_1,i_2}\Big| \geq t\Big\} \leq 2\exp\left(-Ct/A_n\right), \tag{8.39}$$

where $A_n^2 := \sum_{(i_1,i_2)\in \mathbf{i}_2^n} a_{i_1,i_2}^2$. In the next corollary, we leverage Theorem 8.4 and present an alternative tail bound for T_{Rad} that precisely captures the tail bound (8.39) for large t. Note that, unlike i.i.d. Rademacher chaos, T_{Rad} has a non-zero expectation. Hence we construct a tail bound for the chaos statistic centered by $\overline{a} := n_{(2)}^{-1} \sum_{(i_1,i_2)\in \mathbf{i}_2^n} a_{i_1,i_2}$. The proof of the result can be found in Appendix G.15.

Corollary 8.5.1 (Dependent Rademacher chaos)**.** *For every $t > 0$ and some constant $C > 0$, the dependent Rademacher chaos is concentrated as*

$$\mathbb{P}\Big\{ \Big| \sum\nolimits_{(i_1,i_2)\in \mathbf{i}_2^n} \widetilde{\zeta}_{i_1} \widetilde{\zeta}_{i_2} \left(a_{i_1,i_2} - \overline{a}\right) \Big| \geq t \Big\} \leq 2 \exp\left\{ - C \min\left(\frac{t^2}{A_n^2}, \ \frac{t}{A_n} \right) \right\}.$$

The next section studies adaptive tests based on the combinatorial concentration bounds provided in this section.

8.7 Adaptive tests

In this section, we revisit two-sample testing and independence testing for Hölder densities considered in Section 8.4.2 and Section 8.6, respectively. As mentioned earlier, minimax optimality of the multinomial tests for Hölder densities depends on an unknown smoothness parameter (see Proposition 8.3 and Proposition 8.7). The aim of this section is to introduce adaptive permutation tests to this unknown parameter at the expense of an iterated logarithm factor. To this end, we generally follow the Bonferroni-type approach in Ingster (2000) combined with the exponential concentration bounds in Section 8.6.2. Here and hereafter, we restrict to our attention to the nominal level α less than $e^{-1} \approx 0.368$, for which $\log(1/\alpha)$ is larger than $\sqrt{\log(1/\alpha)}$, to simplify our results.

Two-sample testing. Let us start with the two-sample problem. Without loss of generality, assume that $n_1 \leq n_2$ and consider a set of integers such that $\mathsf{K} := \{2^j : j = 1, \ldots, \gamma_{\max}\}$ where

$$\gamma_{\max} := \left\lceil \frac{2}{d} \log_2 \left(\frac{n_1}{\log\log n_1} \right) \right\rceil.$$

For each $\kappa \in \mathsf{K}$, we denote by $\phi_{\kappa,\alpha/\gamma_{\max}} := \mathbb{1}(U_{n_1,n_2} > c_{1-\alpha/\gamma_{\max},n})$, the multinomial two-sample test in Proposition 8.3 with the bin size κ^{-1}. We note that the type I error of an individual test is controlled at $\alpha/\gamma_{\max}$ instead of α. By taking the maximum of the resulting tests, we introduce an adaptive test for two-sample testing as follows:

$$\phi_{\mathrm{adapt}} := \max_{\kappa\in\mathsf{K}} \phi_{\kappa,\alpha/\gamma_{\max}}.$$

This adaptive test does not require knowledge on the smoothness parameter but is still minimax rate optimal up to a small factor of $\log\log n_1$. We describe this result in the following proposition.

Proposition 8.8 (Adaptive two-sample test)**.** *Consider the same problem setting in Proposition 8.3 with an additional assumption that $n_1 \asymp n_2$. For a sufficiently large $C(s, d, L, \alpha, \beta) > 0$, consider ϵ_{n_1,n_2} such that*

$$\epsilon_{n_1,n_2} \geq C(s, d, L, \alpha, \beta) \left(\frac{\log\log n_1}{n_1} \right)^{\frac{2s}{4s+d}}.$$

Then for testing $\mathcal{P}_0 = \{(P_Y, P_Z) \in \mathcal{P}^{(d,s)}_{\text{Hölder}} : f_Y = f_Z\}$ against $\mathcal{P}_1 = \{(P_Y, P_Z) \in \mathcal{P}^{(d,s)}_{\text{Hölder}} : \|f_Y - f_Z\|_{L_2} \geq \epsilon_{n_1,n_2}\}$, the type I and II errors of the adaptive test ϕ_{adapt} are uniformly controlled as in (8.1).

Type I error control of the adaptive test is trivial via the union bound. The proof of the type II error control is an application of Theorem 8.3 and can be found in Appendix G.17. We note that the assumption $n_1 \asymp n_2$ is necessary to apply the concentration result in Theorem 8.3, and it remains an open question whether the same result can be established without $n_1 \asymp n_2$.

Independence testing. Let us now turn to the independence testing problem. Similarly as before, we define a set of integers by $\mathsf{K}^\dagger := \{2^j : j = 1, \ldots, \gamma^*_{\max}\}$ where

$$\gamma^*_{\max} := \left\lceil \frac{2}{d_1 + d_2} \log_2 \left(\frac{n}{\log\log n} \right) \right\rceil.$$

For each $\kappa \in \mathsf{K}^\dagger$, we use the notation $\phi^\dagger_{\kappa,\alpha/\gamma^*_{\max}} := \mathbb{1}(U_n > c_{1-\alpha/\gamma^*_{\max},n})$ to denote the multinomial independence test in Proposition 8.7 with the bin size κ^{-1}. Again we note that the type I error of an individual test is controlled at $\alpha/\gamma^*_{\max}$ instead of α. We then introduce an adaptive test for independence testing by taking the maximum of individual tests as

$$\phi^\dagger_{\text{adapt}} := \max_{\kappa \in \mathsf{K}^\dagger} \phi^\dagger_{\kappa,\alpha/\gamma^*_{\max}}.$$

As in the two-sample case, the adaptive test does not depend on the smoothness parameter. In addition, when densities are smooth enough such that $4s > d_1 + d_2$, the adaptive test is minimax rate optimal up to an iterated logarithm factor as shown in the next proposition.

Proposition 8.9 (Adaptive independence test)**.** *Consider the same problem setting in Proposition 8.6 and suppose that $4s > d_1 + d_2$. For a sufficiently large $C(s, d_1, d_2, L, \alpha, \beta) > 0$, consider ϵ_n such that*

$$\epsilon_n \geq C(s, d_1, d_2, L, \alpha, \beta) \left(\frac{\log\log n}{n} \right)^{\frac{2s}{4s+d_1+d_2}}.$$

Then for testing $\mathcal{P}_0 = \{P_{YZ} \in \mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)} : f_{YZ} = f_Y f_Z\}$ *against* $\mathcal{P}_1 = \{P_{YZ} \in \mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)} : \|f_{YZ} - f_Y f_Z\|_{L_2} \geq \epsilon_n\}$*, the type I and II errors of the resulting permutation test are uniformly controlled as in (8.1).*

The proof of this result relies on Theorem 8.5 and is similar to that of Proposition 8.8. The details can be found in Appendix G.17. The restriction $4s > d_1 + d_2$ is imposed to guarantee that the first term $nt\Lambda_n^{-1}$ is smaller than the second term $nt^{2/3}M_n^{-3/2}$ in the tail bound (8.38) with high probability. Although it seems difficult, we believe that this restriction can be dropped with a more careful analysis. Alternatively one can convert independence testing to two-sample testing via sample-splitting (see Section 8.8.3 for details) and then apply the adaptive two-sample test in Proposition 8.8. The resulting test has the same theoretical guarantee as in Proposition 8.9 without this restriction. However the sample-splitting approach should be considered with caution as it only uses a fraction of the data, which may result in a loss of power in practice.

Remark 8.3 (Comparison to the two moments method)**.** While the exponential inequalities in Section 8.6 lead to the adaptivity at the cost of $\log\log n$ factor, they are limited to degenerate U-statistics and require additional assumptions such as $n_1 \asymp n_2$ and $4s > d_1 + d_2$ to obtain minimax rates. On the other hand, the two moments method is applicable beyond U-statistics and yields minimax rates without these extra assumptions. However we highlight that this generality comes at the cost of $\log n$ factor rather than $\log\log n$ to obtain the same adaptivity results.

8.8 Further applications

In this section, we further investigate the power performance of permutation tests under different problem settings. One specific problem that we focus on is testing for multinomial distributions in the ℓ_1 distance. The ℓ_1 distance has an intuitive interpretation in terms of the total variation distance and has been considered as a metric for multinomial distribution testing (see e.g. Paninski, 2008; Chan et al., 2014; Diakonikolas and Kane, 2016; Balakrishnan and Wasserman, 2019, and also references therein). Unlike the previous work, we approach this problem using the permutation procedure and study its minimax rate optimality in the ℓ_1 distance. We also consider the problem of testing for continuous distributions and demonstrate the performance of the permutation tests based on reproducing kernel-based test statistics in Section 8.8.4 and Section 8.8.5.

8.8.1 Two-sample testing under Poisson sampling with equal sample sizes

Let p_Y and p_Z be multinomial distributions defined on $\mathbb{S}_d$. Suppose that we observe samples from Poisson distributions as $\{Y_{1,k}, \ldots, Y_{n,k}\} \overset{i.i.d.}{\sim} \text{Poisson}\{p_Y(k)\}$ and $\{Z_{1,k}, \ldots, Z_{n,k}\} \overset{i.i.d.}{\sim} \text{Poisson}\{p_Z(k)\}$ for each $k \in \{1, \ldots, d\}$. Assume that all these samples are mutually independent. Let us write $V_k := \sum_{i=1}^n Y_{i,k}$

and $W_k := \sum_{i=1}^n Z_{i,k}$ where V_k and W_k have Poisson distributions with parameters $np_Y(k)$ and $np_Z(k)$, respectively. Under this distributional assumption, Chan et al. (2014) consider a centered chi-square test statistic given by

$$T_{\chi^2} := \sum_{k=1}^{d} \frac{(V_k - W_k)^2 - V_k - W_k}{V_k + W_k} \mathbb{1}(V_k + W_k > 0). \tag{8.40}$$

Based on this statistic, they show that if one rejects the null $H_0 : p_Y = p_Z$ when T_{χ^2} is greater than $C\sqrt{\min\{n, d\}}$ for some constant C, then the resulting test is minimax rate optimal for the class of alternatives determined by the ℓ_1 distance. In particular, the minimax rate is shown to be

$$\epsilon_n^\dagger \asymp \max\left\{\frac{d^{1/2}}{n^{3/4}}, \frac{d^{1/4}}{n^{1/2}}\right\}. \tag{8.41}$$

However, in their test, the choice of C is implicit and based on a loose concentration inequality. Here, by letting $\{X_{i,k}\}_{i=1}^{2n}$ be the pooled samples of $\{Y_{i,k}\}_{i=1}^n$ and $\{Z_{i,k}\}_{i=1}^n$, we instead determine the critical value via the permutation procedure. In this setting the permuted test statistic is

$$T_{\chi^2}^\pi := \sum_{k=1}^{d} \frac{(\sum_{i=1}^n X_{\pi_i,k} - \sum_{i=1}^n X_{\pi_{i+n},k})^2 - V_k - W_k}{V_k + W_k} \mathbb{1}(V_k + W_k > 0).$$

The next theorem shows that the resulting permutation test is also minimax rate optimal.

Theorem 8.6 (Two-sample testing under Poisson sampling)**.** *Consider the distributional setting described above. For a sufficiently large $C > 0$, let us consider a positive sequence ϵ_n such that*

$$\epsilon_n \geq \frac{C}{\beta}\sqrt{\log\left(\frac{1}{\alpha}\right)} \cdot \max\left\{\frac{d^{1/2}}{n^{3/4}}, \frac{d^{1/4}}{n^{1/2}}\right\}.$$

Then for testing $\mathcal{P}_0 = \{(p_Y, p_Z) : p_Y = p_Z\}$ against $\mathcal{P}_1 = \{(p_Y, p_Z) : \|p_Y - p_Z\|_1 \geq \epsilon_n\}$, the type I and II errors of the permutation test based on T_{χ^2} are uniformly controlled as in (8.1).

It is worth noting that $\sqrt{\log(1/\alpha)}$ factor in Theorem 8.6 is a consequence of applying the exponential concentration inequality in Section 8.6. We also note that this logarithmic factor cannot be obtained by the technique used in Chan et al. (2014) which only bounds the mean and variance of the test statistic. On the other hand, the dependency on β may be sub-optimal and may be improved via a more sophisticated analysis. We leave this direction to future work.

8.8.2 Two-sample testing via sample-splitting

Although the chi-square two-sample test in Theorem 8.6 is simple and comes with a theoretical guarantee of minimax optimality, it is only valid in the setting of equal sample sizes. The goal of this subsection is to provide an alternative permutation test via sample-splitting which is minimax rate optimal regardless of the sample size ratio. When the two sample sizes are different, Bhattacharya and Valiant (2015) modify the chi-square statistic (8.40) and propose an optimal test but with the additional assumption that $\epsilon_{n_1,n_2} \geq d^{-1/12}$. Diakonikolas and Kane (2016) remove this extra assumption and introduce another test with the same statistical guarantee. Their test is based on the flattening idea that artificially transforms the probability distributions to be roughly uniform. The same idea is considered in Canonne et al. (2018) for conditional independence testing. Despite their optimality, neither Bhattacharya and Valiant (2015) nor Diakonikolas and Kane (2016) presents a concrete way of choosing the critical value that leads to a level α test. Here we address this issue based on the permutation procedure.

Suppose that we observe $\mathcal{Y}_{2n_1}$ and $\mathcal{Z}_{2n_2}$ samples from two multinomial distributions p_Y and p_Z defined on $\mathbb{S}_d$, respectively. Without loss of generality, we assume that $n_1 \leq n_2$. Let us define $m := \min\{n_2, d\}$ and denote data-dependent weights, computed based on $\{Z_{n_2+1}, \ldots, Z_{n_2+m}\}$, by

$$w_k := \frac{1}{2d} + \frac{1}{2m} \sum_{i=1}^{m} \mathbb{1}(Z_{i+n_2} = k) \quad \text{for } k = 1, \ldots, d.$$

Under the given scenario, we consider the two-sample U-statistic (8.6) defined with the following bivariate function:

$$g_{\text{Multi},w}(x, y) := \sum_{k=1}^{d} w_k^{-1} \mathbb{1}(x = k) \mathbb{1}(y = k). \tag{8.42}$$

We emphasize that the considered U-statistic is evaluated based on the first n_1 observations from each group, i.e. $\mathcal{X}_{2n_1}^{\text{split}} := \{Y_1, \ldots, Y_{n_1}, Z_1, \ldots, Z_{n_1}\}$, which are clearly independent of weights $\{w_1, \ldots, w_d\}$. Let us denote the U-statistic computed in this way by $U_{n_1,n_2}^{\text{split}}$. Let us consider the critical value of a permutation test obtained by permuting the labels within $\mathcal{X}_{2n_1}^{\text{split}}$. Then the resulting permutation test via sample-splitting has the following theoretical guarantee.

Proposition 8.10 (Multinomial two-sample testing in the ℓ_1 distance)**.** *Let $\mathcal{P}_{\text{Multi}}^{(d)}$ be the set of pairs of multinomial distributions defined on $\mathbb{S}_d$. Let $\mathcal{P}_0 = \{(p_Y, p_Z) \in \mathcal{P}_{\text{Multi}}^{(d)} : p_Y = p_Z\}$ and $\mathcal{P}_1(\epsilon_{n_1,n_2}) = \{(p_Y, p_Z) \in \mathcal{P}_{\text{Multi}}^{(d)} : \|p_Y - p_Z\|_1 \geq \epsilon_{n_1,n_2}\}$ where*

$$\epsilon_{n_1,n_2} \geq \frac{C}{\beta^{3/4}} \sqrt{\log\left(\frac{1}{\alpha}\right)} \cdot \max\left\{ \frac{d^{1/2}}{n_1^{1/2} n_2^{1/4}}, \frac{d^{1/4}}{n_1^{1/2}} \right\}, \tag{8.43}$$

for a sufficiently large $C > 0$. Consider the two-sample U-statistic $U^{\mathrm{split}}_{n_1,n_2}$ described above. Then, over the classes $\mathcal{P}_0$ and $\mathcal{P}_1$, the type I and II errors of the resulting permutation test via sample-splitting are uniformly bounded as in (8.1).

Proof Sketch. The proof of this result can be found in Appendix G.20. To sketch the proof, conditional on weights $w_1, \ldots, w_d$, the problem of interest is essentially the same as that of Proposition 8.1. One difference is that $U^{\mathrm{split}}_{n_1,n_2}$ is not an unbiased estimator of $\|p_Y - p_Z\|_1$. However, by noting that $\sum_{k=1}^d w_i = 1$, one can lower bound the expected value in terms of the ℓ_1 distance by Cauchy-Schwarz inequality as

$$\mathbb{E}_P\big[U^{\mathrm{split}}_{n_1,n_2} | w_1, \ldots, w_n\big] = \sum_{k=1}^{d} \frac{\{p_Y(k) - p_Z(k)\}^2}{w_k} \geq \|p_Y - p_Z\|_1^2.$$

The conditional variance can be similarly bounded as in Proposition 8.1 and we use Theorem 8.3 to study the critical value of the permutation test. Finally, we remove the randomness from the weights $w_1, \ldots, w_d$ via Markov's inequality to complete the proof. □

The results of Bhattacharya and Valiant (2015) and Diakonikolas and Kane (2016) show that the minimum separation for ℓ_1-closeness testing satisfies

$$\epsilon^{\dagger}_{n_1,n_2} \asymp \max\left\{ \frac{d^{1/2}}{n_2^{1/4} n_1^{1/2}},\ \frac{d^{1/4}}{n_1^{1/2}} \right\}.$$

This means that the proposed permutation test is minimax rate optimal for multinomial testing in the ℓ_1 distance. On the other hand the procedure depends on sample-splitting which may result in a loss of practical power. Indeed all of the previous approaches (Acharya et al., 2014; Bhattacharya and Valiant, 2015; Diakonikolas and Kane, 2016) also depend on sample-splitting, which leaves the important question as to whether it is possible to obtain the same minimax guarantee without sample-splitting.

8.8.3 Independence testing via sample-splitting

We now turn to independence testing for multinomial distributions in the ℓ_1 distance. To take full advantage of the two-sample test developed in the previous subsection, we follow the idea of Diakonikolas and Kane (2016) in which the independence testing problem is converted into the two-sample problem as follows. Suppose that we observe $\mathcal{X}_{3n}$ samples from a joint multinomial distribution p_{YZ} on $\mathbb{S}_{d_1,d_2}$. We then take the first one-third of the data and denote it by $\widetilde{Y}_n := \{(Y_1, Z_1), \ldots, (Y_n, Z_n)\}$. Using the remaining data, we define another set of samples $\widetilde{Z}_n := \{(Y_{n+1}, Z_{2n+1}), \ldots, (Y_{2n}, Z_{3n})\}$. By construction, it is clear that $\widetilde{Y}_n$ consists of samples from the joint distribution p_{YZ} whereas $\widetilde{Z}_n$ consists of samples from the product distribution $p_Y p_Z$. In other words, we have a fresh dataset $\widetilde{\mathcal{X}}_n := \widetilde{Y}_n \cup \widetilde{Z}_n$ for two-sample testing. It is interesting to mention,

however, that the direct application of the two-sample test in Proposition 8.10 to $\widetilde{\mathcal{X}}_n$ does not guarantee optimality. In particular, by replacing d with $d_1 d_2$ and letting $n_1 = n_2 = n$ in condition (8.43), we see that the permutation test has power when ϵ_{n_1,n_2} is sufficiently larger than $\max\{d_1^{1/2} d_2^{1/2} n^{-3/4}, d_1^{1/4} d_2^{1/4} n^{-1/2}\}$, whereas by assuming $d_1 \le d_2$, the minimum separation for independence testing in the ℓ_1 distance (Diakonikolas and Kane, 2016) is given by

$$\epsilon_n^\dagger \asymp \max\left\{\frac{d_1^{1/4} d_2^{1/2}}{n^{3/4}}, \frac{d_1^{1/4} d_2^{1/4}}{n^{1/2}}\right\}.$$

The main reason is that, unlike the original two-sample problem where two distributions can be arbitrary different, we have further restriction that the marginal distributions of p_{YZ} are the same as those of $p_Y p_Z$. Therefore we need to consider a more refined weight function for independence testing to derive an optimal test. To this end, for each $(k_1, k_2) \in \mathbb{S}_{d_1,d_2}$, we define a product weight by

$$w_{k_1,k_2} := \left[\frac{1}{2d_1} + \frac{1}{2m_1}\sum_{i=1}^{m_1} \mathbb{1}(Y_{3n/2+i} = k_1)\right] \times \left[\frac{1}{2d_2} + \frac{1}{2m_2}\sum_{j=1}^{m_2} \mathbb{1}(Z_{5n/2+i} = k_2)\right],$$

where $m_1 := \min\{n/2, d_1\}$ and $m_2 := \min\{n/2, d_2\}$ and we assume n is even. Notice that by construction, the given product weights are independent of the first half of $\widetilde{\mathcal{X}}_n$, denoted by $\widetilde{\mathcal{X}}_{n/2}^{\text{split}}$. Similarly as before, we use $\widetilde{\mathcal{X}}_{n/2}^{\text{split}}$ to compute the two-sample U-statistic (8.6) defined with the following bivariate function:

$$g^*_{\text{Multi},w}\{(x_1,y_1),(x_2,y_2)\} := \sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2} w_{k_1,k_2}^{-1} \mathbb{1}(x_1 = k_1, y_1 = k_2)\mathbb{1}(x_2 = k_1, y_2 = k_2),$$

and denote the resulting test statistic by $U^{\text{split}*}_{n_1,n_2}$. The critical value is determined by permuting the labels within $\widetilde{\mathcal{X}}_{n/2}^{\text{split}}$ and the resulting test has the following theoretical guarantee.

Proposition 8.11 (Multinomial independence testing in ℓ_1 distance)**.** *Let $\mathcal{P}_{\text{Multi}}^{(d_1,d_2)}$ be the set of multinomial distributions defined on $\mathbb{S}_{d_1,d_2}$. Let $\mathcal{P}_0 = \{p_{YZ} \in \mathcal{P}_{\text{Multi}}^{(d_1,d_2)} : p_{YZ} = p_Y p_Z\}$ and $\mathcal{P}_1(\epsilon_n) = \{p_{YZ} \in \mathcal{P}_{\text{Multi}}^{(d_1,d_2)} : \|p_{YZ} - p_Y p_Z\|_2 \ge \epsilon_n\}$ where*

$$\epsilon_n \ge \frac{C}{\beta^{3/4}}\sqrt{\log\left(\frac{1}{\alpha}\right)} \cdot \max\left\{\frac{d_1^{1/4} d_2^{1/2}}{n^{3/4}}, \frac{d_1^{1/4} d_2^{1/4}}{n^{1/2}}\right\},$$

for a sufficiently large $C > 0$ and $d_1 \le d_2$. Consider the two-sample U-statistic $U^{\text{split}*}_{n_1,n_2}$ described above. Then, over the classes $\mathcal{P}_0$ and $\mathcal{P}_1$, the type I and II errors of the resulting permutation test via sample-splitting are uniformly bounded as in (8.1).

The proof of this result follows similarly as that of Proposition 8.10 with a slight modification due to different kinds of weights. The details are deferred to Appendix G.21. We note again that sample-splitting

is mainly for technical convenience and it might result in a loss of efficiency in practice. An interesting direction of future work is therefore to see whether one can obtain the same minimax guarantee without sample-splitting.

8.8.4 Gaussian MMD

In this subsection we switch gears to continuous distributions and focus on the two-sample U-statistic with a Gaussian kernel. For $x, y \in \mathbb{R}^d$ and $\lambda_1, \ldots, \lambda_d > 0$, the Gaussian kernel is defined by

$$g_{\mathrm{Gau}}(x, y) := K_{\lambda_1,\ldots,\lambda_d,d}(x - y) = \frac{1}{(2\pi)^{d/2}\lambda_1 \cdots \lambda_d} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{d} \frac{(x_i - y_i)^2}{\lambda_i^2} \right\}. \tag{8.44}$$

The two-sample U-statistic defined with this Gaussian kernel is known as the Gaussian maximum mean discrepancy (MMD) statistic due to Gretton et al. (2012) and is also related to the test statistic considered in Anderson et al. (1994). The Gaussian MMD statistic has a nice property that its expectation becomes zero if and only if $P_Y = P_Z$. Given the U-statistic with the Gaussian kernel, we want to find a sufficient condition under which the resulting permutation test has non-trivial power against alternatives determined with respect to the L_2 distance. In detail, by letting f_Y and f_Z be the density functions of P_Y and P_Z with respect to Lebesgue measure, consider the set of paired distributions (P_Y, P_Z) such that the infinity norms of their densities are uniformly bounded, i.e. $\max\{\|f_Y\|_\infty, \|f_Z\|_\infty\} \leq M_{f,d} < \infty$. We denote such a set by $\mathcal{P}_\infty^d$. Then for the class of alternatives $\mathcal{P}_1(\epsilon_{n_1,n_2}) = \{(P_Y, P_Z) \in \mathcal{P}_\infty^d : \|f_Y - f_Z\|_{L_2} \geq \epsilon_{n_1,n_2}\}$, the following proposition gives a sufficient condition on ϵ_{n_1,n_2} under which the permutation-based MMD test has non-trivial power. It is worth noting that a similar result exists in Fromont et al. (2013) where they study the two-sample problem for Poisson processes using a wild bootstrap method. The next proposition differs from their result in three different ways: (1) we consider the usual i.i.d. sampling scheme, (2) we do not assume that n_1 and n_2 are the same and (3) we use the permutation procedure, which is more generally applicable than the wild bootstrap procedure.

Proposition 8.12 (Gaussian MMD)**.** *Consider the permutation test based on the two-sample U-statistic U_{n_1,n_2} with the Gaussian kernel where we assume $\prod_{i=1}^d \lambda_i \leq 1$ and $n_1 \asymp n_2$. For a sufficiently large $C(M_{f,d}, d) > 0$, consider ϵ_{n_1,n_2} such that*

$$\begin{aligned} \epsilon_{n_1,n_2}^2 \;\geq\; & \|(f_Y - f_Z) - (f_Y - f_Z) * K_{\lambda,d}\|_{L_2}^2 \\ & + \frac{C(M_{f,d}, d)}{\beta\sqrt{\lambda_1 \cdots \lambda_d}} \log\left(\frac{1}{\alpha}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right), \end{aligned} \tag{8.45}$$

where $*$ is the convolution operator with respect to Lebesgue measure. Then for testing $\mathcal{P}_0 = \{(P_Y, P_Z) \in \mathcal{P}_\infty^d : f_Y = f_Z\}$ against $\mathcal{P}_1 = \{(P_Y, P_Z) \in \mathcal{P}_\infty^d : \|f_Y - f_Z\|_{L_2} \geq \epsilon_{n_1,n_2}\}$, the type I and II errors of the resulting permutation test are uniformly controlled as in (8.1).

The proof of this result is based on the exponential concentration inequality in Theorem 8.3 and the details are deferred to Appendix G.22. One can remove the assumption that $n_1 \asymp n_2$ using the two moment method in Theorem 8.1 but in this case, the result relies on a polynomial dependence on α. The first term on the right-hand side of condition (8.45) can be interpreted as a bias term, which measures a difference between the L_2 distance and the Gaussian MMD. The second term is related to the variance of the test statistic. We note that there is a certain trade-off between the bias and the variance, depending on the choice of tuning parameters $\{\lambda_i\}_{i=1}^d$. To make the bias term more explicit, we make some regularity conditions on densities, following Fromont et al. (2013) and Meynaoui et al. (2019), and discuss the optimal choice of $\{\lambda_i\}_{i=1}^d$ under each condition.

Example 8.1 (Sobolev ball)**.** For $s, R > 0$, the Sobolev ball $\mathcal{S}_d^s(R)$ is defined as

$$\mathcal{S}_d^s(R) := \left\{ q : \mathbb{R}^d \mapsto \mathbb{R} \middle/ q \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d),\ \int_{\mathbb{R}^d} \|u\|^{2s} |\widehat{q}(u)|^2 du \leq (2\pi)^d R^2 \right\},$$

where $\widehat{q}$ is the Fourier transform of q, i.e. $\widehat{q}(u) := \int_{\mathbb{R}^d} q(x) e^{i\langle x,u\rangle} dx$ and $\langle x, u\rangle$ is the scalar product in $\mathbb{R}^d$. Suppose that $f_Y - f_Z \in \mathcal{S}_d^s(R)$ where $s \in (0, 2]$. Then following Lemma 3 of Meynaoui et al. (2019), it can be seen that the bias term is bounded by

$$\|(f_Y - f_Z) - (f_Y - f_Z) * K_{\lambda,d}\|_{L_2}^2 \leq C(R, s, d) \sum_{k=1}^d \lambda_k^{2s}.$$

Now we further upper bound the right-hand side of condition (8.45) using the above result and then optimize it over $\lambda_1, \ldots, \lambda_d$. This can be done by putting $\lambda_1 = \cdots = \lambda_d = (n_1^{-1} + n_2^{-1})^{2/(4s+d)}$, which in turn yields

$$\epsilon_{n_1,n_2} \geq C(M_{f,d}, R, s, d, \alpha, \beta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{2s}{4s+d}}. \tag{8.46}$$

In other words, Proposition 8.12 holds over the Sobolev ball as long as condition (8.46) is satisfied.

By leveraging the minimax lower bound result in Meynaoui et al. (2019) and the proof of Proposition 8.2, it is straightforward to prove that the minimum separation rate for two-sample testing over the Sobolev ball is $n_1^{-2s/(4s+d)}$ for $n_1 \leq n_2$. This means that the permutation-based MMD test is minimax rate optimal over the Sobolev ball. In the next example, we consider an anisotropic Nikol'skii-Besov ball that can have different regularity conditions over $\mathbb{R}^d$.

Example 8.2 (Nikol'skii-Besov ball)**.** For $\boldsymbol{s} := (s_1, \ldots, s_d) \in (0, \infty)^d$ and $R > 0$, the anisotropic Nikol'skii-Besov ball $\mathcal{N}_{2,d}^{\boldsymbol{s}}(R)$ defined by

$$\mathcal{N}_{2,d}^{\boldsymbol{s}}(R) := \Bigg\{ q : \mathbb{R}^d \mapsto \mathbb{R} \Big/ q \text{ has continuous partial derivatives } D_i^{\lfloor s_i \rfloor} \text{ of order } \lfloor s_i \rfloor \\ \text{with respect to } u_i \text{ and for all } i = 1, \ldots, d,\ u_1, \ldots, u_d, v \in \mathbb{R}, \\ \left\| D_i^{\lfloor s_i \rfloor} q(u_1, \ldots, u_i + v, \ldots, u_d) - D_i^{\lfloor s_i \rfloor} q(u_1, \ldots, u_d) \right\|_{L_2} \leq R|v|^{s_i - \lfloor s_i \rfloor} \Bigg\}.$$

Suppose that $f_Y - f_Z \in \mathcal{N}_{2,d}^{\boldsymbol{s}}(R)$ where $\boldsymbol{s} \in (0, 2]^d$. Then similarly to Lemma 4 of Meynaoui et al. (2019), it can be shown that the bias term is bounded by

$$\left\| (f_Y - f_Z) - (f_Y - f_Z) * K_{\lambda,d} \right\|_{L_2}^2 \leq C(R, \boldsymbol{s}, d) \sum_{k=1}^{d} \lambda_k^{2s_k}.$$

Again we further upper bound the right-hand side of condition (8.45) using the above result and then minimize it over $\lambda_1, \ldots, \lambda_d$. Letting $\eta^{-1} = \sum_{k=1}^{d} s_k^{-1}$, the minimum (up to a constant factor) can be achieved when $\lambda_k = (n_1^{-1} + n_2^{-1})^{2\eta / \{s_k(1+4\eta)\}}$ for $k = 1, \ldots, d$, which yields

$$\epsilon_{n_1,n_2} \geq C(M_{f,d}, R, \boldsymbol{s}, d, \alpha, \beta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{2\eta}{1+4\eta}}. \tag{8.47}$$

Therefore we are guaranteed that Proposition 8.12 holds over the Nikol'skii-Besov ball as long as condition (8.47) is satisfied.

8.8.5 Gaussian HSIC

We now focus on independence testing for continuous distributions. In particular we study the performance of the permutation test using the U-statistic (8.18) defined with Gaussian kernels. For $y_1, y_2 \in \mathbb{R}^{d_1}$, $z_1, z_2 \in \mathbb{R}^{d_2}$ and $\lambda_1, \ldots, \lambda_{d_1}, \gamma_1, \ldots, \gamma_{d_2} > 0$, let us recall the definition of a Gaussian kernel (8.44) and similarly write

$$g_{\mathrm{Gau},Y}(y_1, y_2) := K_{\lambda_1,\ldots,\lambda_{d_1},d_1}(y_1 - y_2) \quad \text{and} \quad g_{\mathrm{Gau},Z}(z_1, z_2) := K_{\gamma_1,\ldots,\gamma_{d_2},d_2}(z_1 - z_2). \tag{8.48}$$

The U-statistic (8.18) defined with these Gaussian kernels is known as the Hilbert–Schmidt independence criterion (HSIC) statistic (Gretton et al., 2005). As in the case of the Gaussian MMD, it is well-known that the expected value of the Gaussian HSIC statistic becomes zero if and only if $P_{YZ} = P_Y P_Z$. Using this property, the resulting test can be consistent against any fixed alternative. Meynaoui et al. (2019) consider the same statistic and study the power of a HSIC-based test over Sobolev and Nikol'skii-Besov balls. It is important to note, however, that the critical value of their test is calculated based on the (theoretical)

null distribution of the test statistic, which is unknown in general. The aim of this subsection is to extend their results to the permutation test that does not require knowledge of the null distribution. To describe the main result, let us write the density functions of P_{YZ} and $P_Y P_Z$ with respect to Lebesgue measure by f_{YZ} and $f_Y f_Z$. As in Section 8.8.4, we use $\mathcal{P}_\infty^{d_1,d_2}$ to denote the set of distributions P_{YZ} whose joint and product densities are uniformly bounded, i.e. $\max\{\|f_{YZ}\|_\infty, \|f_Y f_Z\|_\infty\} \leq M_{f,d_1,d_2} < \infty$. Then the following proposition presents a theoretical guarantee for the permutation-based HSIC test.

Proposition 8.13 (Gaussian HSIC)**.** *Consider the permutation test based on the U-statistic U_n with the Gaussian kernels (8.48) where we assume $\prod_{i=1}^{d_1} \lambda_i \leq 1$ and $\prod_{i=1}^{d_2} \gamma_i \leq 1$. For a sufficiently large $C(M_{f,d_1,d_2}, d_1, d_2) > 0$, consider ϵ_n such that*

$$\begin{aligned} \epsilon_n^2 \;\geq\; & \|(f_{YZ} - f_Y f_Z) - (f_{YZ} - f_Y f_Z) * (K_{\lambda,d_1} K_{\gamma,d_2})\|_{L_2}^2 \\ & + \frac{C(M_{f,d_1,d_2}, d_1, d_2)}{\alpha^{1/2}\beta n \sqrt{\lambda_1 \cdots \lambda_{d_1} \gamma_1 \cdots \gamma_{d_2}}}, \end{aligned} \tag{8.49}$$

where $*$ is the convolution operator with respect to Lebesgue measure. Then for testing $\mathcal{P}_0 = \{P_{YZ} \in \mathcal{P}_\infty^{d_1,d_2} : f_{YZ} = f_Y f_Z\}$ against $\mathcal{P}_1 = \{P_{YZ} \in \mathcal{P}_\infty^{d_1,d_2} : \|f_{YZ} - f_Y f_Z\|_{L_2} \geq \epsilon_n\}$, the type I and II errors of the resulting permutation test are uniformly controlled as in (8.1).

The proof of this result is based on the two moments method in Proposition 8.2. We omit the proof of this result since it is very similar to that of Proposition 8.12 and Theorem 1 of Meynaoui et al. (2019). As before, the first term on the right-hand side of condition (8.49) can be viewed as a bias, which measures a difference between the L_2 distance and the Gaussian HSIC. To make this bias term more tractable, we now consider Sobolev and Nikol'skii-Besov balls and further illustrate Proposition 8.13. The following two examples correspond Corollary 2 and Corollary 3 of Meynaoui et al. (2019) but based on the permutation test.

Example 8.3 (Sobolev ball)**.** Recall the definition of the Sobolev ball from Example 8.1 and assume that $f_{YZ} - f_Y f_Z \in \mathcal{S}_{d_1+d_2}^s(R)$ where $s \in (0, 2]$. Then from Lemma 3 of Meynaoui et al. (2019), the bias term in condition (8.49) can be bounded by

$$\|(f_{YZ} - f_Y f_Z) - (f_{YZ} - f_Y f_Z) * (K_{\lambda,d_1} K_{\gamma,d_2})\|_{L_2}^2 \leq C(R, s, d_1, d_2) \left\{ \sum_{i=1}^{d_1} \lambda_i^{2s} + \sum_{j=1}^{d_2} \gamma_j^{2s} \right\}.$$

For each $i \in \{1, \ldots, d_1\}$ and $j \in \{1, \ldots, d_2\}$, we choose $\lambda_i = \gamma_j = n^{-2/(4s+d_1+d_2)}$ such that the lower bound of ϵ_n in condition (8.49) is minimized. Then by plugging these parameters, it can be seen that Proposition 8.13 holds as long as $\epsilon_n \geq C(M_{f,d_1,d_2}, s, R, d_1, d_2, \alpha, \beta) n^{-\frac{2s}{4s+d_1+d_2}}$. Furthermore this rate matches with the lower bound given in Meynaoui et al. (2019).

Example 8.4 (Nikol'skii-Besov ball)**.** Recall the definition of the Nikol'skii-Besov ball from Example 8.2 and assume that $f_{YZ} - f_Y f_Z \in \mathcal{N}^{\boldsymbol{s}}_{2,d_1+d_2}(R)$ where $\boldsymbol{s} \in (0,2]^{d_1+d_2}$. Then followed by Lemma 4 of Meynaoui et al. (2019), the bias term in condition (8.49) can be bounded by

$$\|(f_{YZ} - f_Y f_Z) - (f_{YZ} - f_Y f_Z) * (K_{\lambda,d_1} K_{\gamma,d_2})\|^2_{L_2} \leq C(R, \boldsymbol{s}, d_1, d_2) \left\{ \sum_{i=1}^{d_1} \lambda_i^{2s_i} + \sum_{j=1}^{d_2} \gamma_j^{2s_{j+d_1}} \right\}.$$

Let us write $\eta^{-1} := \sum_{i=1}^{d_1+d_2} s_i^{-1}$. Then by minimizing the lower bound of ϵ_n in condition (8.49) using the above result with $\lambda_i = n^{-\frac{2\eta}{s_i(1+4\eta)}}$ for $i = 1, \ldots, d_1$ and $\gamma_i = n^{-\frac{2\eta}{s_{i+d_1}(1+4\eta)}}$ for $i = 1, \ldots, d_2$, it can be seen that the same conclusion of Proposition 8.13 holds as long as $\epsilon_n \geq C(M_{f,d_1,d_2}, \boldsymbol{s}, R, d_1, d_2, \alpha, \beta) n^{-\frac{2\eta}{1+4\eta}}$.

From the above two examples, we see that the permutation-based HSIC test has the same power guarantee as the theoretical test considered in Meynaoui et al. (2019). However, our results do not fully recover those in Meynaoui et al. (2019) in terms of α. It remains an open question as to whether Proposition 8.13 continues to hold when $\alpha^{-1/2}$ is replaced by $\log(1/\alpha)$. Alternatively one can employ the sample-splitting idea in Section 8.8.3 and apply the permutation-based MMD test in Proposition 8.12 for independence testing. The result of Proposition 8.12 then guarantees that the MMD test achieves the same rate of the power as the permutation-based HSIC test but it improves the dependency on α in condition (8.49) to a logarithmic factor.

8.9 Simulations

This section provides empirical results to further justify the permutation approach. As emphasized before, the most significant feature of the permutation test is that it tightly controls the type I error rate for any sample size. This is in sharp contrast to non-asymptotic tests based on concentration bounds. The latter tests are typically conservative as they depend on a loose threshold. More seriously it is often the case that this threshold depends on a number of unspecified constants or even unknown parameters which raises the issue of practicality. In the first part of the simulation study, we demonstrate the sensitivity of the latter approach to the choice of constants in terms of type I error control. For this purpose, we focus on the problems of multinomial two-sample and independence testing and the simulation settings are described below.

1. **Two-sample testing.** We consider various power law multinomial distributions under the two-sample null hypothesis. Specifically the probability of each bin is defined to be $p_Y(k) = p_Z(k) \propto k^\gamma$ for $k \in \{1, \ldots, d\}$ and $\gamma \in \{0.2, \ldots, 1.6\}$. We let the sample sizes be $n_1 = n_2 = 50$ and the bin size be $d = 50$. Following Chan et al. (2014) and Diakonikolas and Kane (2016), we use the threshold

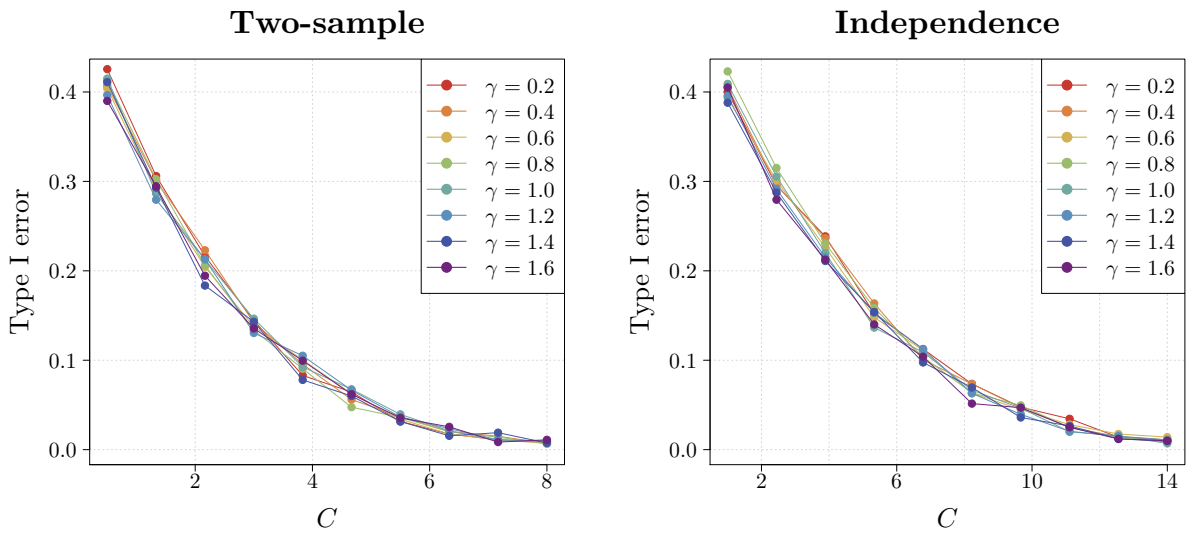

Figure 8.3: Type I error rates of the tests based on concentration bounds by varying constant C in their thresholds. Here we approximated the type I error rates via Monte Carlo simulations under different power law distributions with parameter γ. The results show that the error rates vary considerably depending on the choice of C.

$C\|p_Y\|_2 n_1^{-1}$ for some constant C and reject the null when $U_{n_1,n_2} > C\|p_Y\|_2 n_1^{-1}$ where U_{n_1,n_2} is the U-statistic considered in Proposition 8.1.

2. **Independence testing.** We again consider power law multinomial distributions under the independence null hypothesis. In particular the probability of each bin is defined to be $p_{YZ}(k_1, k_2) = p_Y(k_1)p_Z(k_2) \propto k_1^\gamma k_2^\gamma$ for $k_1, k_2 \in \{1, \ldots, d\}$ and $\gamma \in \{0.2, \ldots, 1.6\}$. We let the sample size be $n = 100$ and the bin sizes be $d_1 = d_2 = 20$. Similarly as before, we use the threshold $C\|p_Y p_Z\|_2 n^{-1}$ for some constant C and reject the null when $U_n > C\|p_Y p_Z\|_2 n^{-1}$ where U_n is the U-statistic considered in Proposition 8.4.

The simulations were repeated 2000 times to approximate the type I error rate of the tests as a function of C. The results are presented in Figure 8.3. One notable aspect of the results is that, in both two-sample and independence cases, the error rates are fairly stable over different null scenarios for each fixed C. However these error rates vary a lot over different C, which clearly shows the sensitivity of the non-asymptotic approach to the choice of C. Furthermore it should be emphasized that both tests are not practical as they depend on unknown parameters $\|p_Y\|_2$ and $\|p_Y p_Z\|_2$, respectively.

It has been demonstrated by several authors (e.g. Hoeffding, 1952) that the permutation distribution of a test statistic mimics the underlying null distribution of the same test statistic in low-dimensional settings. In the next simulation, we provide empirical evidence that the same conclusion still holds in high-dimensional settings. This may further imply that the power of the permutation test approximates that of the theoretical

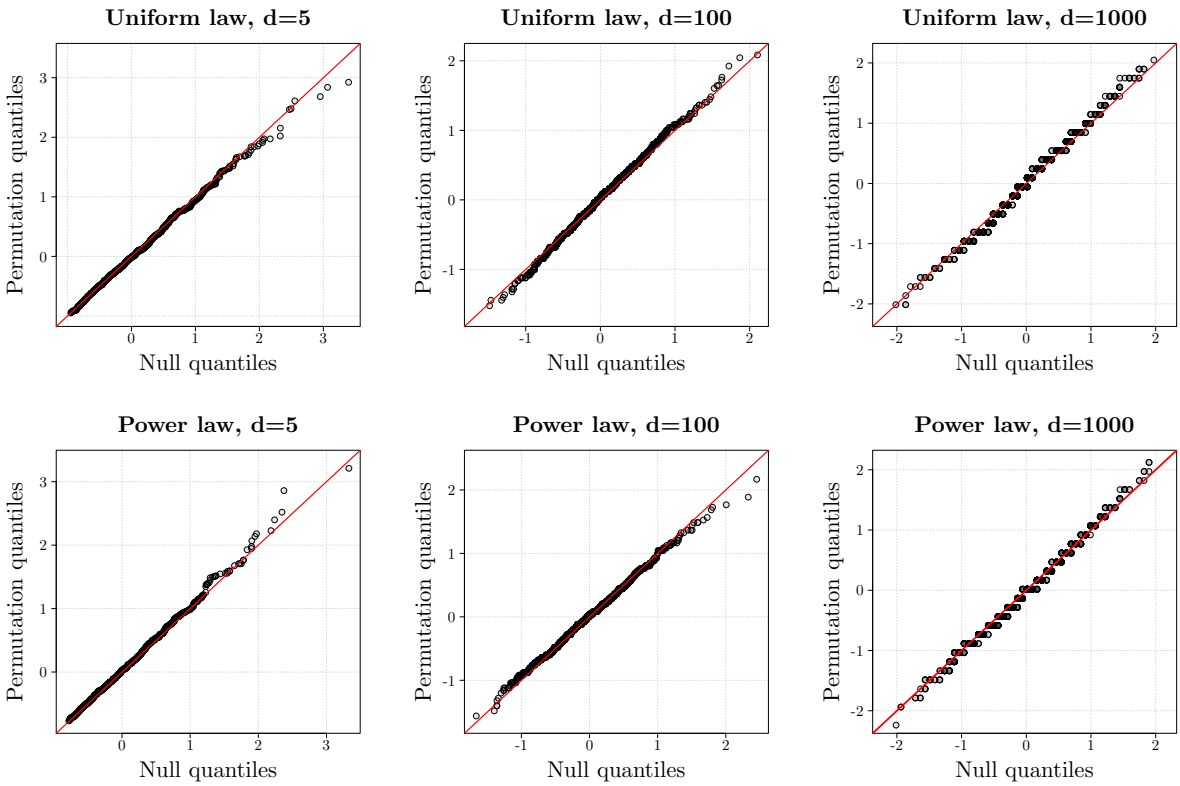

Figure 8.4: Q-Q plots between the null distribution and the permutation distribution of the two-sample U-statistic. The quantiles of the two distributions approximately lie on the straight line $y = x$ in all cases, which demonstrates the similarity of the two distributions. Here we rescaled the test statistic by an appropriate constant for display purpose only.

test based on the null distribution of the test statistic. To illustrate, we focus on the two-sample U-statistic for multinomial testing in Proposition 8.1 and consider two different scenarios as follows.

1. **Uniform law under the null.** We simulate $n_1 = n_2 = 200$ samples from the uniform multinomial distributions under the null such that $p_Y(k) = p_Z(k) = 1/d$ for $k = 1, \ldots, d$ where $d \in \{5, 100, 1000\}$. Conditional on these samples, we compute the permutation distribution of the test statistic. On the other hand, the null distribution of the test statistic is estimated based on $n_1 = n_2 = 200$ samples from the uniform distribution by running a Monte Carlo simulation with 2000 repetitions.

2. **Power law under the alternative.** In order to argue that the power of the permutation test is similar to that of the theoretical test, we need to study the behavior of the permutation distribution under the alternative. For this reason, we simulate $n_1 = 200$ samples from the uniform distribution $p_Y(k) = 1/d$ and $n_2 = 200$ samples from the power law distribution $p_Z(k) \propto k$ for $k = 1, \ldots, d$ where $d \in \{5, 100, 1000\}$. Conditional on these samples, we compute the permutation distribution of

the test statistic. On the other hand, the null distribution of the test statistic is estimated based on $n_1 = n_2 = 200$ samples from the mixture distribution $1/2 \times p_Y + 1/2 \times p_Z$ with 2000 repetitions.

In the simulation study, due to the computational difficulty of considering all possible permutations, we approximated the original permutation distribution using the Monte Carlo method. Nevertheless the difference between the original permutation distribution and its Monte-Carlo counterpart can be made arbitrary small uniformly over the entire real line, which can be shown by using Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956). In our simulations, we randomly sampled 2000 permutations from the entire permutations, based on which we computed the empirical distribution of the permuted test statistic.

We recall from Figure 8.1 that the null distribution of the test statistic changes a lot depending on the size of d. In particular, it tends to be skewed to the right (similar to a χ^2 distribution) when d is small and tends to be symmetric (similar to a normal distribution) when d is large. Also note that the null distribution tends to be more discrete when d is large relative to the sample size. In Figure 8.4, we present the Q-Q plots of the null and (approximate) permutation distributions. It is apparent from the figure that the quantiles of these two distributions approximately lie along the straight line $y = x$ in all the scenarios. In other words, the permutation distribution closely follows the null distribution, regardless of the size of d, from which we conjecture that the null distribution and the permutation distribution might have the same even in high-dimensional settings.

8.10 Discussion

In this work we presented a general framework for analyzing the type II error rate of the permutation test based on the first two moments of the test statistic. We illustrated the utility of the proposed framework in the context of two-sample testing and independence testing in both discrete and continuous cases. In particular, we introduced the permutation tests based on degenerate U-statistics and explored their minimax optimality for multinomial testing as well as density testing. To improve a polynomial dependency on the nominal level α, we developed exponential concentration inequalities for permuted U-statistics based on an idea that links permutations to i.i.d. Bernoulli random variables. The utility of the exponential bounds was highlighted by introducing adaptive tests to unknown parameters and also providing a concentration bound for Rademacher chaos under sampling without replacement.

Our work motivates several lines of future directions. First, while this chapter focused on unconditional independence testing, our results can be extended to conditional independence testing for discrete distributions (e.g. Canonne et al., 2018). When the conditional variable is discrete, one can apply unconditional independence tests within categories and combine them, in a suitable way, to test for conditional independence. When the conditional variable is continuous, however, this strategy does not work. Recently Berrett et al. (2019) proposed a modified permutation procedure for this purpose, and

future research could examine the power of this conditional permutation method, leveraging our results. Second, based on the coupling idea in Section 8.6, further work can be done to develop combinatorial concentration inequalities for other statistics. It would also be interesting to see whether one can obtain tighter concentration bounds, especially for U_n^π in (8.32). Furthermore, improving a polynomial dependency on the type II error rate β is another interesting direction for future research. Finally, we recommend future studies to develop optimal tests for ℓ_1-closeness multinomial testing without sample-splitting.

Chapter 9

Conclusions and future work

In this thesis we have proposed various methods for comparing distributions and investigated their theoretical and empirical properties. One of the main contributions of this thesis is to present a general framework for analyzing permutation procedures and demonstrate their efficacy under nontraditional settings. On a broader perspective, the techniques and results developed throughout this thesis are not limited to the permutation procedure for two-sample and independence testing. Building on the results of this thesis, we hope to address other testing problems and provide a theoretical grounding for the power of data-driven testing procedures under different scenarios. One specific topic that we are currently working on is conditional independence testing. In view of Shah and Peters (2018), there is no valid test for conditional independence that has nontrivial power against any alternative. This remarkable result indicates that it is necessary to make assumptions on the class of null and alternative hypotheses in order to have a meaningful conditional independence test. While this topic has received increasing attention (e.g. Zhang et al., 2012; Candes et al., 2018; Berrett et al., 2019; Neykov et al., 2020), it is still unknown whether there exists a data-driven test that provably controls the type I error rate and also has high power against a broad class of alternatives. We believe that the results of this thesis can serve as building blocks for this direction. Finally, we conclude this chapter with few more concrete topics for future work.

9.1 Limiting behavior and robustness of permutation tests

In classical low-dimensional settings, it is often the case that the permutation test has the same local power as the corresponding asymptotic test while having the advantage of being exact level α under the exchangeability assumption (e.g. Hoeffding, 1952; Robinson, 1973; Romano, 1989; Janssen and Pauls, 2003; Janssen, 2005). However it is largely unknown whether the same asymptotic result can be extended to high-dimensional settings where the sample size increases with other parameters. Leveraging martingale limit theory in

Pauly (2011), we plan to make an initial step towards this question and study the limiting distribution of the permuted test statistic under high-dimensional settings. Currently we expect (with some empirical evidence) that the permutation distribution will have the same limiting law as the corresponding null distribution even in high-dimensions. Leveraging this preliminary result, we aim to prove that the permutation test does not lose any power compared to the asymptotic test even for high-dimensional problems.

Along with this direction, we would like to investigate robustness of the permutation test to the exchangeability assumption. There have been several studies showing that the permutation test based on a properly studentized statistic can be asymptotically exact even when the exchangeability assumption is violated (e.g. Chung and Romano, 2013). We plan to extend the previous low-dimensional results to high-dimensional settings in the context of testing for mean vectors, covariance matrices and regression coefficients.

9.2 Bootstrap approach to high-dimensional inference

The use of a quadratic statistic is common in multivariate data analysis. Classical asymptotic theory shows that this quadratic statistic converges to a weighted sum of independent chi-square random variables where the weights are determined by the covariance structure of the underlying distribution. Over the past decades researchers have identified sharp conditions under which the quadratic statistic further converges to a normal distribution in high-dimensions (e.g. Hall, 1984; Peng and Schick, 2018). One core condition for their results is that the eigenvalues of the covariance are uniformly bounded. However, as observed by Wang and Xu (2019) and others, the limiting distribution of a high-dimensional quadratic statistic is far from normal when the eigenvalue condition is violated. For example, in the extreme case where there exists a single spike in the sequence of eigenvalues, the quadratic statistic essentially has a chi-square distribution with one degree of freedom even in high-dimensional scenarios. One possible remedy for this problem is to estimate the effective number of eigenvalues, which can be used to infer the limiting distribution. However estimating the sequence of eigenvalues is highly non-trivial especially when the dimension is much larger than the sample size. Moreover the variance arising from this extra estimation procedure may result in an adverse effect on overall accuracy.

In the future, we plan to leverage Wang and Xu (2019) and show that the bootstrap approach overcomes the issue. In particular, we plan to generalize the result of Wang and Xu (2019) to a degenerate U-statistic and prove that the bootstrap distribution (and the permutation distribution if applicable) adapts to unknown eigenvalues and converges to the same asymptotic distribution of the U-statistic under various high-dimensional scenarios. As an application of this result, we will try to revisit some high-dimensional testing problems and emphasize the validity of the bootstrap procedure.

Bibliography

Acharya, J., Jafarpour, A., Orlitsky, A., and Suresh, A. T. (2014). Sublinear algorithms for outlier detection and generalized closeness testing. In *2014 IEEE International Symposium on Information Theory*, pages 3200–3204. IEEE. 150, 180

Adamczak, R., Chafaï, D., and Wolff, P. (2016). Circular law for random matrices with exchangeable entries. *Random Structures & Algorithms*, 48(3):454–479. 97

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer. 109

Ahn, J., Marron, J., Muller, K. M., and Chi, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika*, 94(3):760–766. 109

Albers, W., Bickel, P. J., and van Zwet, W. R. (1976). Asymptotic Expansions for the Power of Distribution free Tests in the One-Sample Problem. *The Annals of Statistics*, 4(1):108–156. 151

Albert, M. (2015). *Tests of independence by bootstrap and permutation: an asymptotic and non-asymptotic study. Application to neurosciences.* PhD thesis, Université Nice Sophia Antipolis. 151, 152, 292

Albert, M. (2018). Concentration inequalities for randomly permuted sums. *arXiv preprint arXiv:1805.03579*. 97

Albert, M. (2019). Concentration inequalities for randomly permuted sums. In *High Dimensional Probability VIII*, pages 341–383. Springer. 151, 152, 153, 392, 395

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59(1):19–35. 30

Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88. 149

Anderson, N. H., Hall, P., and Titterington, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54. 27, 58, 182

Anderson, T. W. (1951). Classification by multivariate analysis. *Psychometrika*, 16(1):31–50. 126

Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159. 55, 56, 333

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, volume 3. New York: Wiley-Interscience. 31, 64, 70, 129, 136, 272, 333

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "Goodness of Fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212. 85

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556. 86

Arias-Castro, E., Pelletier, B., and Saligrama, V. (2018). Remember the curse of dimensionality: the case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30(2):448–471. 124, 152, 153, 162, 163, 164, 407, 410, 411, 419, 420, 425

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. 1

Arratia, R., Goldstein, L., and Gordon, L. (1989). Two moments suffice for poisson approximations: the chen-stein method. *The Annals of Probability*, 17(1):9–25. 91, 218, 219

Athreya, K. B. and Lahiri, S. N. (2006). *Measure theory and probability theory*. Springer Science & Business Media. 112, 113, 347, 351

Ayano, T. (2012). Rates of convergence for the k-nearest neighbor estimators with smoother regression functions. *Journal of Statistical Planning and Inference*, 142(9):2530–2536. 35

Babu, G. and Feigelson, E. (2006). Astrostatistics: Goodness-of-fit and all that! In *Astronomical Data Analysis Software and Systems XV*, volume 351, page 127. 1

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6:311–329. 1, 73, 121, 122, 129, 131, 136, 137, 146, 318

Bakshaev, A. (2009). Goodness of fit and homogeneity tests on the basis of n-distances. *Journal of Statistical Planning and Inference*, 139(11):3750–3758. 110

Balakrishnan, N., Voinov, V., and Nikulin, M. S. (2013). *Chi-squared goodness of fit tests with applications*. Academic Press. 8

Balakrishnan, S. and Wasserman, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749. 16

Balakrishnan, S. and Wasserman, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics*, 47(4):1893–1927. 8, 10, 11, 17, 18, 19, 163, 177, 227, 230, 231, 232, 233

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606. 40, 251, 291, 346

Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson approximation*. Clarendon Press Oxford. 233

Bardenet, R. and Maillard, O.-A. (2015). Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385. 398

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206. 45, 56, 58, 104, 110, 113, 114, 159, 330, 351

Baringhaus, L. and Henze, N. (2017). Cramér–von Mises distance: probabilistic interpretation, confidence intervals, and neighbourhood-of-model validation. *Journal of Nonparametric Statistics*, 29(2):167–188. 56

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. (2007). Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 137–144. 123

Bera, A. K., Ghosh, A., and Xiao, Z. (2013). A smooth test for the equality of distributions. *Econometric Theory*, 29(2):419–446. 59

Bercu, B., Delyon, B., and Rio, E. (2015). *Concentration inequalities for sums and martingales*. Springer. 153, 392, 395

Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20(2):1006–1028. 61, 77, 78, 308

Berrett, T. B., Kontoyiannis, I., and Samworth, R. J. (2020). Optimal rates for independence testing via U-statistic permutation tests. *arXiv preprint arXiv:2001.05513*. 152

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2019). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) to appear*. 189, 191

Berry, A. C. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136. 358

Bhat, B. V. (1995). *Theory of U-statistics and its applications*. PhD thesis, Karnatak University. 62, 274, 275, 276, 310, 311, 316, 336

Bhattacharya, B. and Valiant, G. (2015). Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems*, pages 2611–2619. 2, 150, 152, 153, 162, 179, 180

Bhattacharya, B. B. (2018). Two-sample tests based on geometric graphs: Asymptotic distribution and detection thresholds. *arXiv preprint arXiv:1512.00384*. 59, 64, 71, 124

Bhattacharya, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):575–602. 59, 63

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095. 36

Biau, G. and Devroye, L. (2015). *Lectures on the nearest neighbor method*. Springer. 246

Bickel, P. J. and Levina, E. (2004). Some theory for fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010. 130, 131, 137

Bickel, P. J. and Li, B. (2007). Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pages 177–186. 36

Bickel, P. J. and van Zwet, W. R. (1978). Asymptotic Expansions for the Power of Distribution free Tests in the Two-Sample Problem. *The Annals of Statistics*, 6(5):937–1004. 151

Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis*, 123:160–171. 58, 70, 73, 110, 114

Biswas, M., Mukhopadhyay, M., and Ghosh, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika*, 101(4):913–926. 59, 73, 80, 110, 111, 352

Blanchard, G., Lee, G., and Scott, C. (2010). Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11(Nov):2973–3009. 123

Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, 32(2):485–498. 331

Bobkov, S. G. (2004). Concentration of normalized sums and a central limit theorem for noncorrelated random variables. *Annals of Probability*, 32(4):2884–2907. 86, 94, 95

Bogomolny, E., Bohigas, O., and Schmit, C. (2007). Distance matrices and isometric embeddings. *arXiv preprint arXiv:0710.2063*. 329

Bolthausen, E. (1984). An estimate of the remainder in a combinatorial central limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66(3):379–386. 42, 253, 254

Borji, A. (2019). Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding*, 179:41–65. 123

Borwein, J. and Lewis, A. S. (2010). *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media. 358

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press. 97, 171

Bouzebda, S., Keziou, A., and Zari, T. (2011). K-sample problem using strong approximations of empirical copula processes. *Mathematical Methods of Statistics*, 20(1):14–29. 85, 86

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. 43

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media. 36

Bunea, F. and Barbu, A. (2009). Dimension reduction and variable selection in case control studies via regularized likelihood optimization. *Electronic Journal of Statistics*, 3:1257–1287. 30

Burke, M. D. (1979). On the asymptotic power of some k-sample statistics based on the multivariate empirical process. *Journal of Multivariate Analysis*, 9(2):183–205. 85

Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277. 86, 102

Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):349–372. 86, 102

Cai, T. T. and Xia, Y. (2014). High-dimensional sparse MANOVA. *Journal of Multivariate Analysis*, 131:174–196. 86

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:?model-x?knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577. 191

Canonne, C. L. (2015). A survey on distribution testing: Your data is big. but is it blue? In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 22. 2

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. (2018). Testing conditional independence of discrete distributions. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–57. IEEE. 150, 179, 189, 442

Cazáis, F. and Lhéritier, A. (2015). Beyond two-sample-tests: Localizing data discrepancies in high-dimensional spaces. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE. 26

Chakraborty, A. and Chaudhuri, P. (2017). Tests for high-dimensional data based on means, spatial signs and spatial ranks. *The Annals of Statistics*, 45(2):771–799. 58, 73, 74, 75, 78, 326, 327

Chan, S.-O., Diakonikolas, I., Valiant, P., and Valiant, G. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM. 2, 150, 152, 153, 162, 177, 178, 186, 429

Chatterjee, S. (2007). Stein's method for concentration inequalities. *Probability theory and related fields*, 138(1):305–321. 97, 153, 392

Chen, H., Chen, X., and Su, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 113(523):1146–1155. 59, 83

Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112(517):397–409. 59, 116, 117

Chen, L., Dou, W. W., and Qiao, Z. (2013). Ensemble subsampling for imbalanced multivariate two-sample tests. *Journal of the American Statistical Association*, 108(504):1308–1323. 59, 83

Chen, N. F., Shen, W., Campbell, J., and Schwartz, R. (2009). Large-scale analysis of formant frequency estimation variability in conversational telephone speech. In *Tenth Annual Conference of the International Speech Communication Association*. 120

Chen, S. and Pokojovy, M. (2018). Modern and classical k-sample omnibus tests. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(1):e1418. 85

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835. 15, 58, 70, 73, 74, 131, 318

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6):323–329. 43

Chikkagoudar, M. and Bhat, B. V. (2014). Limiting distribution of two-sample degenerate U-statistic under contiguous alternatives and applications. *Journal of Applied Statistical Science*, 22(1–2):127. 64, 262, 266, 267, 276

Childs, D. R. (1967). Reduction of the multivariate normal integral to characteristic form. *Biometrika*, 54(1-2):293–300. 272, 320, 321

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507. 66, 151, 174, 192, 261, 265, 281

Chung, E. and Romano, J. P. (2016a). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *Journal of Statistical Planning and Inference*, 168:97–105. 260

Chung, E. and Romano, J. P. (2016b). Multivariate and multiple permutation tests. *Journal of Econometrics*, 193(1):76–91. 151

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30. 257

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431. 257

Conover, W. (1965). Several k-sample Kolmogorov-Smirnov tests. *The Annals of Mathematical Statistics*, 36(3):1019–1026. 85

Conselice, C. J. (2003). The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1):1. 50

Conselice, C. J. (2014). The evolution of galaxy structure over cosmic time. *Annual Review of Astronomy and Astrophysics*, 52:291–337. 25

Cramér, H. (1928). On the composition of elementary errors. *Skandinavisk Aktuarietidskrift*, 11:141–180. 56

Cui, H. (2002). Average projection type weighted Cramér-von Mises statistics for testing some distributions. *Science in China Series A: Mathematics*, 45(5):562–577. 59

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792. 43

DasGupta, A. (2005). The matching, birthday and the strong birthday problem: a contemporary review. *Journal of Statistical Planning and Inference*, 130(1):377–389. 218

De la Pena, V. and Giné, E. (1999). *Decoupling: from dependence to independence*. Springer Science & Business Media. 173, 174, 423, 425

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media. 35

Diakonikolas, I., Gouleakis, T., Peebles, J., and Price, E. (2016). Collision-based testers are optimal for uniformity and closeness. *arXiv preprint arXiv:1611.03579*. 10, 18

Diakonikolas, I. and Kane, D. M. (2016). A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE. 2, 150, 152, 153, 154, 162, 177, 179, 180, 181, 186

Diakonikolas, I., Kane, D. M., and Nikishkin, V. (2015). Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1841–1854. Society for Industrial and Applied Mathematics. 231

Díaz-Uriarte, R. and De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3. 43

DiCiccio, C. J. and Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, 112(519):1211–1220. 151

Drton, M., Han, F., and Shi, H. (2018). High dimensional independence testing with maxima of rank correlations. *arXiv preprint arXiv:1812.06189*. 86, 89, 90, 91, 335, 340

Dubhashi, D. and Ranjan, D. (1998). Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, 13(2):99–124. 392, 393, 410, 418

Duembgen, L. (1998). Symmetrization and decoupling of combinatorial random elements. *Statistics & Probability Letters*, 39(4):355–361. 153, 169, 173, 174, 394, 423

Duong, T. (2013). Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics*, 25(3):635–645. 23, 27

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669. 156, 189

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press. 150

Eric, M., Bach, F. R., and Harchaoui, Z. (2008). Testing for homogeneity with kernel Fisher discriminant analysis. In *Advances in Neural Information Processing Systems*, pages 609–616. 357

Escanciano, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory*, 22(6):1030–1051. 57, 59, 60, 61, 270, 320

Etzel, J. A., Gazzola, V., and Keysers, C. (2009). An introduction to anatomical ROI-based fMRI classification analysis. *Brain research*, 1282:114–125. 120

Fan, J., Liao, Y., and Yao, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica*, 83(4):1497–1541. 86

Fang, K. W., Kotz, S., and Ng, K. W. (2018). *Symmetric multivariate and related distributions*. Chapman and Hall/CRC. 138

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188. 126

Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics*, 10(1):422–429. 126

Fokianos, K. (2008). Comparing two samples by penalized logistic regression. *Electronic Journal of Statistics*, 2:564–580. 27

Frahm, G. (2004). *Generalized elliptical distributions: theory and applications*. PhD thesis, Universität zu Köln. 138

Freeman, P., Izbicki, R., Lee, A., Newman, J., Conselice, C., Koekemoer, A., Lotz, J., and Mozena, M. (2013). New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society*, 434(1):282–295. 50, 52

Friedman, J. (2004). On multivariate goodness-of-fit and two-sample testing. Technical report, Stanford Linear Accelerator Center, Menlo Park, CA (US). 123

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning*. Springer, New York. 28, 29, 45

Friedman, J. H. (2003). On multivariate goodness of fit and two sample testing. *eConf*, 30908(SLAC-PUB-10325):311–313. 26

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717. 59, 70, 110, 113, 123

Fromont, M., Laurent, B., and Reynaud-Bouret, P. (2013). The two-sample problem for poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *The Annals of Statistics*, 41(3):1431–1461. 152, 158, 182, 183, 292, 398, 400, 429

Fukumizu, K., Gretton, A., Lanckriet, G. R., Schölkopf, B., and Sriperumbudur, B. K. (2009). Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in Neural Information Processing Systems*, pages 1750–1758. 123

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, pages 489–496. 88

Gagnon-Bartsch, J. and Shem-Tov, Y. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics*, 13(3):1464–1483. 26, 43, 122, 123

Giri, N. and Kiefer, J. (1964). Local and asymptotic minimax properties of multivariate tests. *The Annals of Mathematical Statistics*, 35(1):21–35. 127, 136

Giri, N., Kiefer, J., and Stein, C. (1963). Minimax Character of Hotelling's T^2 Test in the Simplest Case. *The Annals of Mathematical Statistics*, 34(4):1524–1535. 127, 136

Golland, P. and Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 330–341. Springer. 120

Gómez, E., Gómez-Villegas, M. A., and Marín, J. M. (2003). A survey on continuous elliptical vector distributions. *Revista matemática complutense*, 16(1):345–361. 138

González-Manteiga, W. and Cao, R. (1993). Testing the hypothesis of a general linear model using nonparametric regression estimation. *Test*, 2(1-2):161–188. 26

González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, 22(3):361–411. 26

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media. 2

Gregory, G. G. (1977). Large sample theory for U-statistics and tests of fit. *The Annals of Statistics*, 5(1):110–123. 266

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A Kernel Method for the Two-Sample Problem. In *Advances in Neural Information Processing Systems*, pages 513–520. 86, 87

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773. 44, 45, 58, 80, 86, 87, 88, 95, 98, 102, 104, 114, 123, 154, 159, 182, 357

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer. 154, 165, 184

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media. 35, 39, 41, 247, 248

Haberman, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association*, 83(402):555–560. 8, 11

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, 14(1):1–16. 192, 223

Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields*, 89(4):447–455. 88

Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic press. 223

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444. 73, 109, 111

Hamza, M. and Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75(8):629–643. 43

Han, F., Chen, S., and Liu, H. (2017). Distribution-free tests of independence in high dimensions. *Biometrika*, 104(4):813–828. 86, 91, 102

Harchaoui, Z., Bach, F., Cappe, O., and Moulines, E. (2013). Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Processing Magazine*, 30(4):87–97. 58

Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 21(4):1926–1947. 26

Hart, J. (2013). *Nonparametric smoothing and lack-of-fit tests*. Springer Science & Business Media. 26

He, H. Y., Basu, K., Zhao, Q., and Owen, A. B. (2019). Permutation p-value approximation via generalized Stolarsky invariance. *The Annals of Statistics*, 47(1):583–611. 94

Hediger, S., Michel, L., and Näf, J. (2019). On the use of random forest for two-sample testing. *arXiv preprint arXiv:1903.06287*. 26, 43, 122, 123, 357

Hemerik, J. and Goeman, J. (2018a). Exact testing with random permutations. *TEST*, 27(4):811–825. 156

Hemerik, J. and Goeman, J. J. (2018b). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):137–155. 142, 386

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783. 58, 110, 123

Hettmansperger, T. P., Möttönen, J., and Oja, H. (1998). Affine invariant multivariate rank tests for several samples. *Statistica Sinica*, 8:785–800. 58

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802. 49

Hodara, P. and Reynaud-Bouret, P. (2019). Exponential inequality for chaos based on sampling without replacement. *Statistics & Probability Letters*, 146:65–69. 174

Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, 19(4):546–557. 330, 331

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, 23(2):169–192. 66, 149, 151, 156, 187, 191, 265

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30. 153, 169, 170, 395, 425

Holst, L. (1972). Asymptotic normality and efficiency for certain goodness-of-fit tests. *Biometrika*, 59(1):137–145. 9

Hotelling, H. (1931). The generalization of student's ratio. *Annals of Mathematical Statistics*, 2(3):360–378. 125

Hu, J. and Bai, Z. (2016). A review of 20 years of naive tests of significance for high-dimensional mean vectors and covariance matrices. *Science China Mathematics*, 59(12):2281–2300. 23, 74, 124, 125

Hušková, M. and Meintanis, S. G. (2008). Tests for the multivariate k-sample problem based on the empirical characteristic function. *Journal of Nonparametric Statistics*, 20(3):263–277. 85, 86, 105

Ingster, Y. I. (1987). Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337. 40, 156, 163, 167, 169, 407, 419

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Mathematical Methods of Statistics*, 2(2):85–114. 407

Ingster, Y. I. (2000). Adaptive chi-square tests. *Journal of Mathematical Sciences*, 99(2):1110–1119. 164, 175, 425

Janssen, A. (2005). Resampling student's t-type statistics. *Annals of the Institute of Statistical Mathematics*, 57(3):507–529. 151, 191

Janssen, A. and Pauls, T. (2003). How do bootstrap and permutation tests work? *The Annals of Statistics*, 31(3):768–806. 151, 191

Jeng, X. J., Cai, T. T., and Li, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association*, 105(491):1156–1166. 86

Jiang, B., Ye, C., and Liu, J. S. (2015). Nonparametric k-sample tests via dynamic slicing. *Journal of the American Statistical Association*, 110(510):642–653. 85

Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *The Annals of Statistics*, 11(1):286–295. 392, 393

Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Transactions on Information Theory*, 58(2):708–720. 59

Kariya, T. (1981). A robustness property of Hotelling's T^2-test. *The Annals of Statistics*, 9(1):211–214. 125, 136

Keziou, A. and Leoni-Aubin, S. (2005). Test of homogeneity in semiparametric two-sample density ratio models. *Comptes Rendus Mathématique*, 340(12):905–910. 27

Kiefer, J. (1959). K-sample analogues of the Kolmogorov-Smirnov and Cramér-von. Mises tests. *The Annals of Mathematical Statistics*, 30(2):420–447. 85

Kim, I. (2019). Comparing a large number of multivariate distributions. *arXiv preprint arXiv:1904.05741*. 85, 152

Kim, I. (2020). Multinomial goodness-of-fit based on u-statistics: High-dimensional asymptotic and minimax optimality. *Journal of Statistical Planning and Inference*, 205:74–91. 7

Kim, I., Balakrishnan, S., and Wasserman, L. (2020a). Minimax optimality of permutation tests. *arXiv preprint arXiv:2003.13208*. 149

Kim, I., Balakrishnan, S., and Wasserman, L. (2020b). Robust multivariate nonparametric tests via projection-averaging. *arXiv preprint arXiv:1803.00715v3 (accepted to the Annals of Statistics)*. 151, 152

Kim, I., Lee, A. B., Lei, J., et al. (2019a). Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305. 23, 151, 152

Kim, I., Ramdas, A., Singh, A., and Wasserman, L. (2019b). Classification accuracy as a proxy for two sample testing. *arXiv preprint arXiv:1602.02210v2*. 152

Kirch, C. and Steinebach, J. (2006). Permutation principles for the change analysis of stochastic processes under strong invariance. *Journal of computational and applied mathematics*, 186(1):64–88. 149

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J., and Shmulevich, I. (2009). Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):i161–i168. 94

Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36. 107

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737. 35, 40

Kpotufe, S. and Garg, V. (2013). Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems*, pages 3075–3083. 35

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50. 80

Lee, J. (1990). *U-statistics: Theory and Practice*. CRC Press. 12, 62, 63, 89, 90, 159, 160, 165, 166, 265, 277, 402, 412

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, 22(2):165–179. 271, 274, 275

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media. 8, 32, 63, 64, 67, 93, 94, 155, 156, 241, 267, 278, 281, 439

Lemeshko, B. Y. and Veretelnikova, I. V. (2018). On Some New K-Samples Tests for Testing the Homogeneity of Distribution Laws. In *2018 XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, pages 153–157. IEEE. 85

Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika*, 105(3):529–546. 348

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940. 70

Li, T. and Yuan, M. (2019). On the Optimality of Gaussian Kernel Based Nonparametric Tests against Smooth Alternatives. *arXiv preprint arXiv:1909.03302*. 104

Lin, Z. and Lu, C. (1997). *Limit theory for mixing dependent random variables*, volume 378. Springer Science & Business Media. 112

Liu, R. Y. (2006). *Data depth: robust multivariate analysis, computational geometry, and applications*, volume 72. American Mathematical Society. 58

Liu, W. and Li, Y. Q. (2020). Sign-based Test for Mean Vector in High-dimensional and Sparse Settings. *Acta Mathematica Sinica, English Series*, 36(1):93–108. 86

Liu, Y., Li, C.-L., and Póczos, B. (2018). Classifier two-sample test for video anomaly detections. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK*, page 71. 120

Liu, Z. and Modarres, R. (2011). A triangle test for equality of distribution functions in high dimensions. *Journal of Nonparametric Statistics*, 23(3):605–615. 59

Lopez-Paz, D. and Oquab, M. (2016). Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*. 26, 59, 122, 123

Lotz, J. M., Primack, J., and Madau, P. (2004). A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, 128(1):163. 50, 52

Luschgy, H. (1982). Minimax character of the two-sample χ^2-test. *Statistica Neerlandica*, 36(3):129–134. 121, 128, 136, 359

Marriott, P., Sabolova, R., Van Bever, G., and Critchley, F. (2015). Geometry of goodness-of-fit testing in high dimensional low sample size modelling. In *International Conference on Networked Geometric Science of Information*, pages 569–576. Springer. 8, 11

Martínez-Camblor, P., De Una-Alvarez, J., and Corral, N. (2008). k-Sample test based on the common area of kernel density estimators. *Journal of Statistical Planning and Inference*, 138(12):4006–4020. 85

Massart, P. (1986). Rates of convergence in the central limit theorem for empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 22, pages 381–423. 397

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283. 103, 345

Mathai, A. M., Provost, S. B., and Hayakawa, T. (2012). *Bilinear forms and zonal polynomials*. Springer Science & Business Media. 381

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188. 98

Meynaoui, A., Laurent, B., Albert, M., and Marrel, A. (2019). Aggregated test of independence based on HSIC measures. *arXiv preprint arXiv:1902.06441*. 154, 183, 184, 185, 186, 446

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K., Hu, Y.-H., Larsen, J., Wilson, E., and Douglas, S. (1999). Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48. 357

Mondal, P. K., Biswas, M., and Ghosh, A. K. (2015). On high dimensional two-sample tests based on nearest neighbors. *Journal of Multivariate Analysis*, 141:168–178. 47, 58, 73, 80, 111, 116, 117

Monhor, D. (2013). Inequalities for correlated bivariate normal distribution function. *Probability in the Engineering and Informational Sciences*, 27(1):115–123. 271

Morris, C. (1975). Central limit theorems for multinomial sums. *The Annals of Statistics*, 3(1):165–188. 9

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2016). Kernel mean embedding of distributions: A review and beyonds. *arXiv preprint arXiv:1605.09522*. 87

Mukhopadhyay, S. and Wang, K. (2018). A Nonparametric Approach to High-dimensional K-sample Comparison Problem. *arXiv preprint arXiv:1810.01724*. 59, 85, 86, 105

Neykov, M., Balakrishnan, S., and Wasserman, L. (2020). Minimax optimal conditional independence testing. *arXiv preprint arXiv:2001.03039*. 191

Oja, H. (2010). *Multivariate nonparametric methods with R: an approach based on spatial signs and ranks*. Springer Science & Business Media. 58

Oja, H. and Randles, R. H. (2004). Multivariate nonparametric tests. *Statistical Science*, 19(4):598–605. 58

Ojala, M. and Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(Jun):1833–1863. 26, 149

Olivetti, E., Greiner, S., and Avesani, P. (2012). Induction in neuroscience with classification: issues and solutions. In *Machine Learning and Interpretation in Neuroimaging*, pages 42–50. Springer. 120

Olivetti, E., Greiner, S., and Avesani, P. (2015). Statistical independence for the evaluation of classifier-based diagnosis. *Brain Informatics*, 2(1):13–19. 26

Pan, W., Tian, Y., Wang, X., and Zhang, H. (2018). Ball Divergence: Nonparametric two sample test. *The Annals of Statistics*, 46(3):1109–1137. 58, 80

Paninski, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755. 177

Park, P. J., Manjourides, J., Bonetti, M., and Pagano, M. (2009). A permutation test for determining significance of clusters with applications to spatial and gene expression data. *Computational statistics & data analysis*, 53(12):4290–4300. 149

Pauly, M. (2011). Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics*, 5:41–52. 192

Pauly, M., Brunner, E., and Konietschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):461–473. 151

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175. 1, 8

Peng, H. and Schick, A. (2018). Asymptotic normality of quadratic forms with random vectors of increasing dimension. *Journal of Multivariate Analysis*, 164:22–39. 192

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 45(1):S199–S209. 120

Pesarin, F. (2001). *Multivariate permutation tests: with applications in biostatistics*. Wiley, New York. 57

Pesarin, F. and Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons. 2, 149, 155

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Science & Business Media. 150

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411. 30

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, pages 609–618. 28

Quessy, J.-F. and Éthier, F. (2012). Cramér–von Mises and characteristic function tests for the two and k-sample problems with dependent data. *Computational Statistics & Data Analysis*, 56(6):2097–2111. 85

Ramdas, A., Reddi, S. J., Poczos, B., Singh, A., and Wasserman, L. (2015). Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *arXiv preprint arXiv:1508.00655*. 45, 73, 74

Ramdas, A., Singh, A., and Wasserman, L. (2016). Classification accuracy as a proxy for two sample testing. *arXiv preprint arXiv:1602.02210*. 26, 30

Raudys, Š. and Young, D. M. (2004). Results in statistical discriminant analysis: A review of the former soviet union literature. *Journal of Multivariate Analysis*, 89(1):1–35. 128

Read, T. R. and Cressie, N. A. (2012). *Goodness-of-fit statistics for discrete multivariate data*. Springer Science & Business Media. 1, 8, 9, 11

Rempała, G. A. and Wesołowski, J. (2016). Double asymptotics for the chi-square statistic. *Statistics & Probability Letters*, 119:317–325. 9, 11

Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons. 358

Rizzo, M. L. and Székely, G. J. (2010). Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055. 86, 104

Robinson, J. (1973). The large-sample power of permutation tests for randomization models. *The Annals of Statistics*, 1(2):291–296. 151, 191

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, 17(1):141–159. 151, 191

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692. 151

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108. 156

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530. 59, 80, 110, 124

Rosenblatt, J., Gilron, R., and Mukamel, R. (2016). Better-than-chance classification for signal detection. *arXiv preprint arXiv:1608.08873*. 26, 30, 33, 123, 145

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9. 422

Salaevskii, O. (1971). Minimax Character of Hotelling's T^2 Test. I. In *Investigations in Classical Problems of Probability Theory and Mathematical Statistics*, pages 74–101. Springer. 127, 136

Sarkar, S. and Ghosh, A. K. (2018). On some high-dimensional two-sample tests based on averages of inter-point distances. *Stat*, 7(1):e187. 109

Sarkar, S. and Ghosh, A. K. (2019). On perfect clustering of high dimension, low sample size data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15. 109

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806. 58, 70, 110, 113, 123

Scholz, F. W. and Stephens, M. A. (1987). K-sample Anderson–Darling tests. *Journal of the American Statistical Association*, 82(399):918–924. 85

Scott, A. J. and Wild, C. (2001). Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96(1):3–27. 30

Scott, C. and Nowak, R. (2005). A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(8):3806–3819. 123

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291. 58, 76, 90

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons. 336

Shah, R. D. and Peters, J. (2018). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics (to appear)*. 191

Simaika, J. (1941). On an optimum property of two important statistical tests. *Biometrika*, 32(1):70–80. 136

Siroker, D. and Koomen, P. (2013). *A/B testing: The most powerful way to turn clicks into customers*. John Wiley & Sons. 1

Slepian, D. (1962). The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2):463–501. 272

Snyder, G. F., Torrey, P., Lotz, J. M., Genel, S., McBride, C. K., Vogelsberger, M., Pillepich, A., Nelson, D., Sales, L. V., and Sijacki, D. (2015). Galaxy morphology and star formation in the illustris simulation at z= 0. *Monthly Notices of the Royal Astronomical Society*, 454(2):1886–1908. 51

Sosthene, A., Balogoun, K., Martial Nkiet, G., and Ogouyandjou, C. (2018). Kernel based method for the k-sample problem. *arXiv preprint arXiv:1812.00100*. 85, 86

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410. 88

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386–402. 121, 136, 138, 143

Srivastava, M. S., Katayama, S., and Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114:349–358. 131

Steck, G. P. (1957). Limit theorems for conditional distributions. 9

Stelzer, J., Chen, Y., and Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *Neuroimage*, 65:69–82. 120

Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. (2011). Least-squares two-sample test. *Neural Networks*, 24(7):735–751. 27

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2016). Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*. 1

Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5(16.10). 45, 56, 58, 104, 110, 114, 159

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272. 56

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794. 165

Thas, O. (2010). *Comparing distributions*. Springer. 1, 23, 55, 85

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. (2017). Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048. 103, 345

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats*. Springer Series in Statistics. Springer, New York. 35, 163, 269, 291

Tumanyan, S. (1954). On the asymptotic distribution of the chi-square criterion. In *Dokl. Akad. Nauk. SSSR*, volume 94, pages 1011–1012. 9

Tumanyan, S. K. (1956). Asymptotic distribution of the χ^2 criterion when the number of observations and number of groups increase simultaneously. *Theory of Probability & Its Applications*, 1(1):117–131. 9

Valiant, G. and Valiant, P. (2017). An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455. 8, 11, 17, 18, 231, 232

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645. 36

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press. 135, 364

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press. 90, 171, 338, 339, 422

Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*. 36

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press. 307

Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *The Annals of Mathematical Statistics*, 15(2):145–162. 126

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162. 59

Wang, C. and Carroll, R. (1993). On robust estimation in logistic case-control studies. *Biometrika*, 80(1):237–241. 30

Wang, L., Peng, B., and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, 110(512):1658–1669. 78

Wang, R. and Xu, X. (2019). A feasible high dimensional randomization test for the mean vector. *Journal of Statistical Planning and Inference*, 199:160–178. 192

Wang, S. and Carroll, R. J. (1999). High-order accurate methods for retrospective sampling problems. *Biometrika*, 86(4):881–897. 30

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science & Business Media. 256

Weihrather, G. (1993). Testing a linear regression model against nonparametric alternatives. *Metrika*, 40(1):367–379. 26

Wyłupek, G. (2010). Data-driven k-sample tests. *Technometrics*, 52(1):107–123. 85

Xiao, J., Wang, R., Teng, G., and Hu, Y. (2014). A transfer learning based classifier ensemble model for customer credit scoring. In *2014 Seventh International Joint Conference on Computational Sciences and Optimization*, pages 64–68. IEEE. 120

Xiao, J., Xiao, Y., Huang, A., Liu, D., and Wang, S. (2015). Feature-selection-based dynamic transfer ensemble model for customer churn prediction. *Knowledge and information systems*, 43(1):29–51. 120

Xu, W., Hou, Y., Hung, Y., and Zou, Y. (2013). A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing*, 93(1):261–276. 272, 320

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599. 40

Yu, K., Martin, R., Rothman, N., Zheng, T., and Lan, Q. (2007). Two-sample Comparison Based on Prediction Error, with Applications to Candidate Gene Association Studies. *Annals of human genetics*, 71(1):107–118. 120

Zaitsev, A. Y. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of SN Bernstein's inequality conditions. *Probability theory and related fields*, 74(4):535–566. 337, 338

Zelnik-Manor, L. and Perona, P. (2005). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, pages 1601–1608. 52, 257

Zelterman, D. (1986). The log-likelihood ratio for sparse multinomial mixtures. *Statistics & Probability Letters*, 4(2):95–99. 9, 11

Zelterman, D. (1987). Goodness-of-fit tests for large sparse multinomial distributions. *Journal of the American Statistical Association*, 82(398):624–629. 9, 11

Zhan, D. and Hart, J. (2014). Testing equality of a large number of densities. *Biometrika*, 101(2):449–464. 86

Zhang, C. and Dette, H. (2004). A power comparison between nonparametric regression tests. *Statistics & Probability Letters*, 66(3):289–301. 26

Zhang, J. and Wu, Y. (2007). k-Sample tests based on the likelihood ratio. *Computational Statistics & Data Analysis*, 51(9):4682–4691. 85

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*. 191

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75(2):263–289. 26

Zhou, W.-X., Zheng, C., and Zhang, Z. (2017). Two-sample smooth tests for the equality of distributions. *Bernoulli*, 23(2):951–989. 59

Zhou, Y.-H., Marron, J., and Wright, F. A. (2018). Eigenvalue significance testing for genetic association. *Biometrics*, 74(2):439–447. 149

Zhu, C.-Z., Zang, Y.-F., Cao, Q.-J., Yan, C.-G., He, Y., Jiang, T.-Z., Sui, M.-Q., and Wang, Y.-F. (2008). Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. *Neuroimage*, 40(1):110–120. 120

Zhu, L., Xu, K., Li, R., and Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika*, 104(4):829–843. 57, 59, 330

Zhu, L.-X., Fang, K.-T., and Bhatti, M. I. (1997). On estimated projection pursuit-type Cramér–von Mises Statistics. *Journal of Multivariate Analysis*, 63(1):1–14. 59, 60

Zollanvari, A., Braga-Neto, U. M., and Dougherty, E. R. (2011). Analytic study of performance of error estimators for linear discriminant analysis. *IEEE Transactions on Signal Processing*, 59(9):4238–4255. 130

Zolotarev, V. M. (1961). Concerning a certain probability problem. *Theory of Probability & Its Applications*, 6(2):201–204. 340, 342

Appendix

Appendix A

Appendix for Chapter 2

A.1 Proofs

A.1.1 Proof of Lemma 2.0.1

We will provide a more general result by considering an arbitrary positive diagonal matrix $A = \text{diag}\{a_1, \ldots, a_d\}$ in the kernel. In other words, we will show the following holds:

$$\sum_{j=1}^{d} a_j \left(Y_j - n\pi_{0,j}\right)^2 = \sum_{1\le i\le n} \sum_{1\le j\le n} (\mathbf{X}_i - \pi_0)^\top A(\mathbf{X}_j - \pi_0). \tag{A.1}$$

Then the result follows by setting $a_i = (n\pi_{0,i})^{-1}$.

First, we decompose the left hand side of (A.1) into the three parts:

$$\sum_{j=1}^{d} a_j \left(Y_j - n\pi_{0,j}\right)^2 = \underbrace{\sum_{j=1}^{d} a_j Y_j^2}_{(i)} - \underbrace{2n \sum_{j=1}^{d} a_j Y_j \pi_{0,j}}_{(ii)} + \underbrace{n^2 \sum_{j=1}^{d} a_j \pi_{0,j}^2}_{(iii)}$$

and treat them separately.

Part (i). Recall that $Y_j = \sum_{i=1}^{n} I(X_{i,j} = 1)$, and thus

$$\begin{aligned}\sum_{j=1}^{d} a_j \left(\sum_{i=1}^{n} I(X_{i,j} = 1)\right)^2 &= \sum_{j=1}^{d} a_j \left[\sum_{i=1}^{n} I(X_{i,j} = 1) + 2\sum_{i<i'} I(X_{i,j} = 1) I(X_{i',j} = 1)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{d} a_j I(X_{i,j} = 1) + 2\sum_{i<i'}\sum_{j=1}^{d} a_j I(X_{i,j} = 1) I(X_{i',j} = 1)\end{aligned}$$

$$= \sum_{i=1}^{n} \mathbf{X}_i^\top A \mathbf{X}_i + 2 \sum_{i<i'} \mathbf{X}_i^\top A \mathbf{X}_{i'} = \sum_{i=1}^{n} \sum_{i'=1}^{n} \mathbf{X}_i^\top A \mathbf{X}_{i'}.$$

Part (ii). Similar to the first part,

$$\begin{aligned} 2n \sum_{j=1}^{d} a_j Y_j \pi_{0,j} &= 2n \sum_{j=1}^{d} a_j \left(\sum_{i=1}^{n} I(X_{i,j} = 1) \right) \pi_{0,j} \\ &= 2n \sum_{i=1}^{n} \sum_{j=1}^{d} a_j I(X_{i,j} = 1) \pi_{0,j} = 2n \sum_{i=1}^{n} \mathbf{X}_i^\top A \pi_0 \\ &= \sum_{i=1}^{n} \sum_{i'=1}^{n} \left(\mathbf{X}_i^\top A \pi_0 + \mathbf{X}_{i'}^\top A \pi_0 \right). \end{aligned}$$

Part (iii). The last part is straightforward,

$$n^2 \sum_{j=1}^{d} a_j \pi_{0,j}^2 = \sum_{i=1}^{n} \sum_{i'=1}^{n} \pi_0^\top A \pi_0.$$

Combining the three parts, we can get the desired result.

A.1.2 Proof of Theorem 2.1

The proof is mainly based on Theorem 1 of Arratia et al. (1989). We describe it in Theorem A.1. Before we get to the main proof, we provide several lemmas.

Recall the following decomposition of U_I:

$$U_I = \binom{n}{2}^{-1} \underbrace{\sum_{1 \le i < j \le n} \mathbf{X}_i^\top \mathbf{X}_j}_{W} - \frac{2}{n} \sum_{i=1}^{n} \left(\mathbf{X}_i^\top \pi_0 + \pi_0^\top \pi_0 \right).$$

Suppose $\binom{n}{2} \pi_0^\top \pi_0 \to \eta$ as $n, d \to \infty$. As preliminary results, we are interested in the conditions that result in

$$W = \sum_{1 \le i < j \le n} \mathbf{X}_i^\top \mathbf{X}_j \xrightarrow{d} \text{Pois}(\eta) \quad \text{where } \eta \in (0, \infty).$$

Note that W is the sum of locally dependent indicator variables. The limiting distribution of W has been studied under the name of the birthday problem (see e.g., DasGupta, 2005). Let Z be a Poisson random

variable with $\mathbb{E}[W] = \mathbb{E}[Z]$. Here, our interest is in the total variation distance between W and Z, i.e.

$$d_{TV}(W, Z) = 2 \sup_{A \in \mathbb{Z}^+} \Big| \mathbb{P}(W \in A) - \mathbb{P}(Z \in A) \Big|.$$

In order to bound the total variation distance, we employ Theorem 1 of Arratia et al. (1989):

Theorem A.1 (Theorem 1 of Arratia et al. (1989))**.** *Let $\mathcal{I}$ be an arbitrary index set, and for $\alpha \in \mathcal{I}$, let X_α be a Bernoulli random variable with $p_\alpha = \mathbb{P}(X_\alpha = 1) > 0$. Let $K = \sum_{\alpha \in \mathcal{I}} X_\alpha$ and $\mathbb{E}[K] = \sum_{\alpha \in \mathcal{I}} p_\alpha \in (0, \infty)$. For each $\alpha \in \mathcal{I}$, suppose we have chosen $B_\alpha \in \mathcal{I}$ with $\alpha \in B_\alpha$. Define*

$$b_1 = \sum_{\alpha \in \mathcal{I}} \sum_{\beta \in B_\alpha} p_\alpha p_\beta, \quad b_2 = \sum_{\alpha \in \mathcal{I}} \sum_{\alpha \neq \beta \in B_\alpha} \mathbb{E}[X_\alpha X_\beta], \quad \text{and} \quad b_3 = \sum_{\alpha \in \mathcal{I}} \mathbb{E} \Big| \mathbb{E} \Big\{ X_\alpha - p_\alpha \Big| \sum_{\beta \in \mathcal{I} - B_\alpha} X_\beta \Big\} \Big|.$$

Let Z be a Poisson random variable with $\mathbb{E}[K] = \mathbb{E}[Z] = \eta$. Then

$$d_{TV}(K, Z) \leq 2 \left[(b_1 + b_2) \frac{1 - e^{-\eta}}{\eta} + b_3 (1 \wedge 1.4 \eta^{-1/2}) \right]. \tag{A.2}$$

As a corollary of Theorem A.1, we present the total variation distance between W and the Poisson random variable Z.

Corollary A.1.1. *Let Z be a Poisson random variable with $\mathbb{E}[Z] = \mathbb{E}[W] = \binom{n}{2} \pi^\top \pi = \eta_{n,d}$. Then*

$$d_{TV}(W, Z) \;\leq\; 2n^3 \frac{1 - e^{-\eta_{n,d}}}{\eta_{n,d}} \left[\sum_{i=1}^d \pi_i^3 + \left\{ \sum_{i=1}^d \pi_i^2 \right\}^2 \right].$$

Proof. Denote $p_1 = \mathbb{P}\left(\mathbf{X}_1^\top \mathbf{X}_2 = 1 \right)$ and $p_2 = \mathbb{P}\left(\mathbf{X}_1^\top \mathbf{X}_2 \mathbf{X}_1^\top \mathbf{X}_3 = 1 \right)$. In addition, let $\mathcal{I}$ be a set of all indices (i, j) where $1 \leq i < j \leq n$ so that $|\mathcal{I}| = \binom{n}{2}$ and $\mathcal{B}_i$ be a collection of $\mathbf{X}_j^\top \mathbf{X}_k$ that is dependent on $\mathbf{X}_i$. Here $|\mathcal{B}_i| = \binom{n}{2} - \binom{n-2}{2} = 2n - 3$ for all i.

Thus we have

$$b_1 = p_1^2 |\mathcal{I}| |\mathcal{B}_i| = \left\{ \sum_{i=1}^d \pi_i^2 \right\}^2 \binom{n}{2} (2n - 3) \leq \left\{ \sum_{i=1}^d \pi_i^2 \right\}^2 n^3,$$

$$b_2 = p_2 |\mathcal{I}| \left(|\mathcal{B}_i| - 1 \right) = \sum_{i=1}^d \pi_i^3 \binom{n}{2} (2n - 4) = \sum_{i=1}^d \pi_i^3 n(n-1)(n-2) \leq \sum_{i=1}^d \pi_i^3 n^3.$$

and $b_3 = 0$ by the construction of $\mathcal{B}_i$. Then the proof is complete by applying Theorem A.1. □

So far, we have investigated the asymptotic results of W. Now, let us turn our attention to U_I. Recall that U_I and W are related as

$$\binom{n}{2} U_I = W - (n-1)\sum_{i=1}^{n} \mathbf{X}_i^\top \pi_0 + \binom{n}{2} \pi_0^\top \pi_0.$$

Hence, in order to attain the Poisson approximation for $\binom{n}{2} U_I$, one might need to control the last two terms properly. Note that

$$\begin{aligned}
\mathbb{E}\left[(n-1)\sum_{i=1}^{n} \mathbf{X}_i^\top \pi_0\right] &= 2\binom{n}{2} \pi^\top \pi_0 = 2\eta_2 + o(1), \\
\mathrm{Var}\left[(n-1)\sum_{i=1}^{n} \mathbf{X}_i^\top \pi_0\right] &= (n-1)^2 n \left[\mathbb{E}\left[\mathbf{X}_1^\top \pi_0 \mathbf{X}_1^\top \pi_0\right] - \left(\mathbb{E}\left[\mathbf{X}_1^\top \pi_0\right]\right)^2\right] \\
&\le n^3 \left(\sum_{i=1}^{d} \pi_i \pi_{0,i}^2 - \left(\sum_{i=1}^{d} \pi_i \pi_{0,i}\right)^2\right).
\end{aligned}$$

Hence, under (P.2) and (P.3), Chebyshev's inequality yields

$$(n-1)\sum_{i=1}^{n} \mathbf{X}_i^\top \pi_0 \xrightarrow{p} 2\eta_2. \tag{A.3}$$

We finish the proof by applying Slutsky's theorem.

A.1.3 Proof of Corollary 2.1.1 and 2.1.2

These results are direct applications of Theorem 2.1; hence we omit the proof.

A.1.4 Variance of U_A

In the next lemma, we calculate the closed-form of the variance of U_A.

Lemma A.1.1 (Variance of U_A)**.** *Let A be a symmetric positive definite matrix and Σ be* $\mathrm{diag}(\pi) - \pi\pi^\top$. *Then*

$$\mathrm{Var}\,(U_A) = \binom{n}{2}^{-1} \Big\{ \mathrm{tr}\{(A\Sigma)^2\} + 2(n-1)(\pi - \pi_0)^\top A\Sigma A(\pi - \pi_0) \Big\}. \tag{A.4}$$

Therefore,

$$\mathrm{Var}\,(U_{\pi_0}) = \binom{n}{2}^{-1} \Big\{ \mathrm{tr}\{(D_{\pi_0}\Sigma)^2\} + 2(n-1)(\pi - \pi_0)^\top D_{\pi_0}\Sigma D_{\pi_0}(\pi - \pi_0) \Big\} \quad \textit{and}$$

$$\mathrm{Var}\left(U_I\right) = \binom{n}{2}^{-1}\Big\{\mathrm{tr}\left(\Sigma^2\right) + 2(n-1)(\pi-\pi_0)^\top\Sigma(\pi-\pi_0)\Big\},$$

where D_{π_0} is defined in (2.3).

Proof. Without loss of generality, we assume $A = I$. Otherwise, define $\mathbf{X}_1^* = A^{1/2}\mathbf{X}_1$, $\mathbf{X}_2^* = A^{1/2}\mathbf{X}_2$ and $\pi_0^* = A^{1/2}\pi_0$ so that $h_A(\mathbf{X}_1, \mathbf{X}_2) = \left(\mathbf{X}_1^* - \pi_0^*\right)^\top\left(\mathbf{X}_2^* - \pi_0^*\right)$, and proceed similarly. Note that the variance of double summations becomes

$$\begin{aligned}&\mathrm{Var}\left(\sum_{1\le i<j\le n} h_I(\mathbf{X}_i, \mathbf{X}_j)\right)\\ &= \binom{n}{2}\mathrm{Var}\left(h_I(\mathbf{X}_1, \mathbf{X}_2)\right) + 2(n-2)\mathrm{Cov}\left(h_I(\mathbf{X}_1, \mathbf{X}_2), h_I(\mathbf{X}_1, \mathbf{X}_3)\right).\end{aligned} \tag{A.5}$$

We treat the variance and the covariance separately. First, calculate the variance of the kernel:

$$\begin{aligned}\mathrm{Var}\left(h_I(\mathbf{X}_1, \mathbf{X}_2)\right) &= \mathbb{E}\left[h_I^2(\mathbf{X}_1, \mathbf{X}_2)\right] - \left\{\mathbb{E}\left[h_I(\mathbf{X}_1, \mathbf{X}_2)\right]\right\}^2\\ &= \mathbb{E}\left[h_I^2(\mathbf{X}_1, \mathbf{X}_2)\right] - ||\pi - \pi_0||_2^2.\end{aligned}$$

The expected value can be decomposed into:

$$\begin{aligned}\mathbb{E}\left[h_I^2(\mathbf{X}_1, \mathbf{X}_2)\right] &= \mathbb{E}\left[\left\{\left(\mathbf{X}_1 - \pi_0\right)^\top\left(\mathbf{X}_2 - \pi_0\right)\right\}^2\right]\\ &= \mathbb{E}\left[\left\{\left(\mathbf{X}_1 - \pi + \pi - \pi_0\right)^\top\left(\mathbf{X}_2 - \pi + \pi - \pi_0\right)\right\}^2\right]\\ &= \mathbb{E}\left[\left\{(\mathbf{X}_1 - \pi)^\top(\mathbf{X}_2 - \pi)\right\}^2\right] + 2\mathbb{E}\left[\left\{(\mathbf{X}_1 - \pi)^\top(\pi - \pi_0)\right\}^2\right] + ||\pi - \pi_0||_2^2.\end{aligned}$$

The first term can be simplified as

$$\begin{aligned}\mathbb{E}\left[\left\{(\mathbf{X}_1 - \pi)^\top(\mathbf{X}_2 - \pi)\right\}^2\right] &= \mathbb{E}\left[\mathrm{tr}\left\{(\mathbf{X}_1 - \pi)^\top(\mathbf{X}_2 - \pi)(\mathbf{X}_2 - \pi)^\top(\mathbf{X}_1 - \pi)\right\}\right]\\ &= \mathbb{E}\left[\mathrm{tr}\left\{(\mathbf{X}_1 - \pi)(\mathbf{X}_1 - \pi)^\top(\mathbf{X}_2 - \pi)(\mathbf{X}_2 - \pi)^\top\right\}\right]\\ &= \mathrm{tr}\Big\{\mathbb{E}\left[(\mathbf{X}_1 - \pi)(\mathbf{X}_1 - \pi)^\top(\mathbf{X}_2 - \pi)(\mathbf{X}_2 - \pi)^\top\right]\Big\}\\ &= \mathrm{tr}\Big\{\mathbb{E}\left[(\mathbf{X}_1 - \pi)(\mathbf{X}_1 - \pi)^\top\right]\mathbb{E}\left[(\mathbf{X}_2 - \pi)(\mathbf{X}_2 - \pi)^\top\right]\Big\}\\ &= \mathrm{tr}\left(\Sigma^2\right).\end{aligned}$$

On the other hand, the second term becomes

$$\begin{aligned}\mathbb{E}\left[\left\{(\mathbf{X}_1-\pi)^\top(\pi-\pi_0)\right\}^2\right] &= \mathbb{E}\left[(\pi-\pi_0)^\top(\mathbf{X}_1-\pi)(\mathbf{X}_1-\pi)^\top(\pi-\pi_0)\right] \\ &= (\pi-\pi_0)^\top\Sigma(\pi-\pi_0).\end{aligned}$$

Hence, the variance of the kernel can be calculated by

$$\mathrm{Var}\left(h_I(\mathbf{X}_1,\mathbf{X}_2)\right) = \mathrm{tr}\left(\Sigma^2\right) + 2(\pi-\pi_0)^\top\Sigma(\pi-\pi_0). \tag{A.6}$$

Next, turn our attention to the covariance:

$$\begin{aligned}\mathrm{Cov}\left(h_I(\mathbf{X}_1,\mathbf{X}_2), h_I(\mathbf{X}_1,\mathbf{X}_3)\right) &= \mathbb{E}\left[h_I(\mathbf{X}_1,\mathbf{X}_2)h_I(\mathbf{X}_1,\mathbf{X}_3)\right] - \mathbb{E}\left[h_I(\mathbf{X}_1,\mathbf{X}_2)\right]\mathbb{E}\left[h_I(\mathbf{X}_1,\mathbf{X}_3)\right] \\ &= \mathbb{E}\left[h_I(\mathbf{X}_1,\mathbf{X}_2)h_I(\mathbf{X}_1,\mathbf{X}_3)\right] - ||\pi-\pi_0||_2^4. \end{aligned} \tag{A.7}$$

Note that

$$\begin{aligned}\mathbb{E}\left[h_I(\mathbf{X}_1,\mathbf{X}_2)h_I(\mathbf{X}_1,\mathbf{X}_3)\right] &= \mathbb{E}\left[(\mathbf{X}_1-\pi_0)^\top(\mathbf{X}_2-\pi_0)(\mathbf{X}_1-\pi_0)^\top(\mathbf{X}_3-\pi_0)\right] \\ &= \mathbb{E}\left[(A_1+B)^\top(A_2+B)(A_1+B)^\top(A_3+B)\right], \end{aligned} \tag{A.8}$$

where $A_i = \mathbf{X}_i - \pi$ for $i = 1, 2, 3$, and $B = \pi - \pi_0$. In fact, (A.8) is equivalent to

$$\mathbb{E}\left[A_1^\top B A_1^\top B\right] + B^\top B B^\top B = (\pi-\pi_0)^\top\Sigma(\pi-\pi_0) + ||\pi-\pi_0||_2^4, \tag{A.9}$$

due to

$$\begin{aligned}\mathbb{E}\left[A_1^\top A_2 A_3^\top B\right] &= \mathbb{E}\left[(\mathbf{X}_1-\pi)^\top(\mathbf{X}_2-\pi)(\mathbf{X}_3-\pi)^\top(\pi-\pi_0)\right] \\ &= \mathbb{E}\left[(\mathbf{X}_1-\pi)^\top\left\{\mathbb{E}\left[(\mathbf{X}_2-\pi)(\mathbf{X}_3-\pi)^\top\middle|\mathbf{X}_1\right]\right\}(\pi-\pi_0)\right] \\ &= \mathbb{E}\left[(\mathbf{X}_1-\pi)^\top\left\{\mathbb{E}\left[(\mathbf{X}_2-\pi)\middle|\mathbf{X}_1\right]\mathbb{E}\left[(\mathbf{X}_3-\pi)^\top\middle|\mathbf{X}_1\right]\right\}(\pi-\pi_0)\right] \\ &= 0.\end{aligned}$$

In the same way, we can see the other terms become zero. Then, we get a simple form of the covariance from (A.7) and (A.9):

$$\mathrm{Cov}\left(h_I(\mathbf{X}_1,\mathbf{X}_2), h_I(\mathbf{X}_1,\mathbf{X}_3)\right) = (\pi-\pi_0)^\top\Sigma(\pi-\pi_0). \tag{A.10}$$

We finish the proof by multiplying $\binom{n}{2}^{-2}$ to (A.5) together with (A.6) and (A.10). □

A.1.5 Proof of Theorem 2.2

The proof is based on Corollary 3.1 of Hall and Heyde (1980). Under the null, U_A is a degenerate centered U-statistic, which satisfies $\mathbb{E}\left[h_A(\mathbf{X}_1, \mathbf{X}_2)\right] = 0$ and $\mathbb{E}\left[h_A(\mathbf{X}_1, \mathbf{X}_2)\middle|\mathbf{X}_2\right] = 0$. We follow the similar proof steps in Theorem 1 of Hall (1984), but we adapt the argument to obtain the convergence result for the uniform null in Corollary 2.2.1.

First, we define the filtration $\mathcal{F}_k = \sigma\left(\mathbf{X}_1, \ldots, \mathbf{X}_k\right)$, and let

$$Y_j = \sum_{i=1}^{j-1} h_A(\mathbf{X}_i, \mathbf{X}_j) \quad \text{and} \quad \mathcal{S}_k = \sum_{j=2}^{k} Y_j,$$

for $2 \leq k \leq n$. It is easy to check that $\{(\mathcal{S}_k, \mathcal{F}_k)\}$ is a square integrable martingale sequence with zero mean as

$$\mathbb{E}\left[\mathcal{S}_j\right] = 0 \quad \text{and} \quad \mathbb{E}\left[\mathcal{S}_i\middle|\mathcal{F}_j\right] = \mathcal{S}_j + \sum_{k=j+1}^{i} \mathbb{E}\left[Y_k\middle|\mathcal{F}_j\right] = \mathcal{S}_j$$

for any $i \geq j$. Denote the variance of $\sum_{i<j} h_A(\mathbf{X}_i, \mathbf{X}_j)$ by $s_n^2 = \binom{n}{2}\mathrm{tr}\{(A\Sigma)^2\}$. Then, according to Corollary 3.1 of Hall and Heyde (1980), it is enough to show that the following two conditions are satisfied under the given assumptions:

(C.1) $s_n^{-2} \sum_{i=2}^{n} \mathbb{E}\left[Y_i^2 I\left(|Y_i| > \epsilon s_n\right)\right] \;\to\; 0.$

(C.2) $s_n^{-2} \sum_{i=2}^{n} \mathbb{E}\left[Y_i^2\middle|\mathcal{F}_{i-1}\right] \xrightarrow{p} 1.$

Let us first verify the first condition (C.1). Since $|Y_i| > \epsilon s_n$ implies

$$Y_i^2 = \frac{|Y_i|^{2+\delta}}{|Y_i|^{\delta}} \leq \frac{|Y_i|^{2+\delta}}{(\epsilon s_n)^{\delta}},$$

for any $\epsilon, \delta > 0$, we have

$$s_n^{-2} \sum_{i=2}^{n} \mathbb{E}\left[Y_i^2 I\left(|Y_i| > \epsilon s_n\right)\right] \leq \epsilon^{-\delta} s_n^{-2-\delta} \sum_{i=2}^{n} \mathbb{E}\left[|Y_i|^{2+\delta}\right].$$

By choosing $\delta = 2$, we will show that $s_n^{-4}\sum_{i=2}^n \mathbb{E}\big[Y_i^4\big] \to 0$ to verify (C.1). From the fact that, for any distinct (i_1, i_2, i_3, i_4) or for any combination (i_1, i_2, i_3, i_4) where only one of them is different,

$$\mathbb{E}\left[h_A(\mathbf{X}, \mathbf{X}_{i_1})h_A(\mathbf{X}, \mathbf{X}_{i_2})h_A(\mathbf{X}, \mathbf{X}_{i_3})h_A(\mathbf{X}, \mathbf{X}_{i_4})\right] = 0,$$

we can see that

$$\begin{aligned}\mathbb{E}\left[Y_i^4\right] &= \sum_{i_1,i_2,i_3,i_4=1}^{i-1} \mathbb{E}\left[h_A(\mathbf{X}_i, \mathbf{X}_{i_1})h_A(\mathbf{X}_i, \mathbf{X}_{i_2})h_A(\mathbf{X}_i, \mathbf{X}_{i_3})h_A(\mathbf{X}_i, \mathbf{X}_{i_4})\right] \\ &= (i-1)\mathbb{E}\left[\left\{h_A(\mathbf{X}_1, \mathbf{X}_2)\right\}^4\right] + 3(i-1)(i-2)\mathbb{E}\left[\left\{h_A(\mathbf{X}_1, \mathbf{X}_2)\right\}^2\left\{h_A(\mathbf{X}_1, \mathbf{X}_3)\right\}^2\right].\end{aligned}$$

Hence, we have

$$\sum_{i=2}^n \mathbb{E}\big[Y_i^4\big] = \frac{n(n-1)}{2}\mathbb{E}\left[\left\{h_A(\mathbf{X}_1, \mathbf{X}_2)\right\}^4\right] + n(n-1)(n-2)\mathbb{E}\left[\left\{h_A(\mathbf{X}_1, \mathbf{X}_2)\right\}^2\left\{h_A(\mathbf{X}_1, \mathbf{X}_3)\right\}^2\right].$$

From the second assumption in (2.8), it is easy to see $s_n^{-4}\sum_{i=2}^n \mathbb{E}\big[Y_i^4\big] \to 0$, which verifies (C.1).

Now, we prove that (C.2) holds under the given conditions, that is to show

$$\frac{2}{n(n-1)\mathrm{tr}\{(A\Sigma)^2\}}\sum_{i=2}^n \mathbb{E}\left[Y_i^2\big|\mathcal{F}_{i-1}\right] \xrightarrow{p} 1.$$

First, we can see from $\mathbb{E}\left[h_A(\mathbf{X}_1, \mathbf{X}_2)h_A(\mathbf{X}_1, \mathbf{X}_3)\right] = 0$ and $\mathbb{E}\left[h_A^2(\mathbf{X}_1, \mathbf{X}_2)\right] = \mathrm{tr}\{(A\Sigma)^2\}$, that

$$\begin{aligned}\sum_{i=2}^n \mathbb{E}\left[Y_i^2\right] &= \sum_{i=2}^n \sum_{j_1,j_2=1}^{i-1} \mathbb{E}\left[h_A(\mathbf{X}_i, \mathbf{X}_{j_1})h_A(\mathbf{X}_i, \mathbf{X}_{j_2})\right] \\ &= \sum_{i=2}^n (i-1)\mathbb{E}\left[h_A(\mathbf{X}_i, \mathbf{X}_1)\right] = \frac{n(n-1)}{2}\mathrm{tr}\{(A\Sigma)^2\}.\end{aligned}$$

Therefore, it is sufficient to prove

$$\frac{4}{n^2(n-1)^2\mathrm{tr}\{(A\Sigma)^2\}^2}\sum_{i_1,i_2=2}^n \mathrm{Cov}\left(\mathbb{E}\left[Y_{i_1}^2\big|\mathcal{F}_{i_1-1}\right], \mathbb{E}\left[Y_{i_2}^2\big|\mathcal{F}_{i_2-1}\right]\right) \to 0.$$

Let us define

$$\begin{aligned}G_A(\mathbf{X}_i, \mathbf{X}_j) &= \mathbb{E}\left[h_A(\mathbf{X}_i, \mathbf{X}_k)h_A(\mathbf{X}_j, \mathbf{X}_k)\big|\sigma(\mathbf{X}_i, \mathbf{X}_j)\right] \\ &= (\mathbf{X}_i - \pi_0)^\top A\Sigma_0 A(\mathbf{X}_j - \pi_0), \qquad \text{(A.11)}\end{aligned}$$

so that

$$\mathbb{E}\left[Y_i^2|\mathcal{F}_{i-1}\right] = \sum_{j_1,j_2=1}^{i-1} \mathbb{E}\left[h_A(\mathbf{X}_i,\mathbf{X}_{j_1})h_A(\mathbf{X}_i,\mathbf{X}_{j_2})\middle|\mathcal{F}_{i-1}\right] = \sum_{j_1,j_2=1}^{i-1} G_A(\mathbf{X}_{j_1},\mathbf{X}_{j_2}).$$

Then the covariance becomes

$$\mathrm{Cov}\left(\mathbb{E}\left[Y_{i_1}^2\middle|\mathcal{F}_{i_1-1}\right],\mathbb{E}\left[Y_{i_2}^2\middle|\mathcal{F}_{i_2-1}\right]\right) = \sum_{j_1,j_2=1}^{i_1-1}\sum_{j_1',j_2'=1}^{i_2-1} \mathrm{Cov}\left(G_A(\mathbf{X}_{j_1},\mathbf{X}_{j_2}),G_A(\mathbf{X}_{j_1'},\mathbf{X}_{j_2'})\right).$$

Note that for $j_1 \le j_2$ and $j_1' \le j_2'$,

$$\mathrm{Cov}\left(G_A(\mathbf{X}_{j_1},\mathbf{X}_{j_2}),G_A(\mathbf{X}_{j_1'},\mathbf{X}_{j_2'})\right) = \begin{cases} \mathrm{Var}\left(G_A(\mathbf{X}_1,\mathbf{X}_1)\right) & \text{if } j_1=j_2=j_1'=j_2' \\ \mathbb{E}\left[G_A(\mathbf{X}_1,\mathbf{X}_2)^2\right] & \text{if } j_1=j_1'\neq j_2=j_2' \\ 0 & \text{otherwise} \end{cases}.$$

Hence, if $i_1 \ge i_2$,

$$\mathrm{Cov}\left(\mathbb{E}\left[Y_{i_1}^2\middle|\mathcal{F}_{i_1-1}\right],\mathbb{E}\left[Y_{i_2}^2\middle|\mathcal{F}_{i_2-1}\right]\right) = (i_2-1)\mathrm{Var}\left(G_A(\mathbf{X}_1,\mathbf{X}_1)\right) + 2(i_2-1)(i_2-2)\mathbb{E}\left[G_A(\mathbf{X}_1,\mathbf{X}_2)^2\right]$$

and the sum of the covariance becomes

$$\sum_{i_1,i_2=2}^{n} \mathrm{Cov}\left(\mathbb{E}\left[Y_{i_1}^2\middle|\mathcal{F}_{i_1-1}\right],\mathbb{E}\left[Y_{i_2}^2\middle|\mathcal{F}_{i_2-1}\right]\right) \le C_1\left\{n^3\mathrm{Var}\left(G_A(\mathbf{X}_1,\mathbf{X}_1)\right) + n^4\mathbb{E}\left[G_A(\mathbf{X}_1,\mathbf{X}_2)^2\right]\right\},$$

where C_1 is a constant independent on n. Using (A.11), we have

$$\begin{aligned}
\mathbb{E}\left[G_A(\mathbf{X}_1,\mathbf{X}_2)^2\right] &= \mathbb{E}\left[\left\{(\mathbf{X}_1-\pi_0)^\top A\Sigma_0 A(\mathbf{X}_2-\pi_0)\right\}^2\right] = \mathrm{tr}\left\{(A\Sigma_0)^4\right\}, \\
\mathrm{Var}\left(G_A(\mathbf{X}_1,\mathbf{X}_1)\right) &= \mathbb{E}\left[\left\{h_A(\mathbf{X}_1,\mathbf{X}_2)\right\}^2\left\{h_A(\mathbf{X}_1,\mathbf{X}_3)\right\}^2\right] - \left\{\mathbb{E}\left[\left\{h_A(\mathbf{X}_1,\mathbf{X}_2)\right\}^2\right]\right\}^2 \\
&\le \mathbb{E}\left[\left\{h_A(\mathbf{X}_1,\mathbf{X}_2)\right\}^2\left\{h_A(\mathbf{X}_1,\mathbf{X}_3)\right\}^2\right].
\end{aligned}$$

Now, under the given conditions, we bound

$$\begin{aligned}
&\frac{4}{n^2(n-1)^2\mathrm{tr}\{(A\Sigma)^2\}^2}\sum_{i_1,i_2=2}^{n}\mathrm{Cov}\left(\mathbb{E}\left[Y_{i_1}^2\middle|\mathcal{F}_{i_1-1}\right],\mathbb{E}\left[Y_{i_2}^2\middle|\mathcal{F}_{i_2-1}\right]\right) \\
\le\ & C_2\left(\frac{\mathrm{tr}\{(A\Sigma)^4\} + n^{-1}\mathbb{E}\left[\left\{h_A(\mathbf{X}_1,\mathbf{X}_2)\right\}^2\left\{h_A(\mathbf{X}_1,\mathbf{X}_3)\right\}^2\right]}{\mathrm{tr}\{(A\Sigma)^2\}^2}\right) \ \to\ 0
\end{aligned}$$

where C_2 is a constant independent on n. This completes the proof.

A.1.6 Proof of Corollary 2.2.1

Note that the variance of U_I of the uniform null distribution is

$$\binom{n}{2}^{-1} \mathrm{tr}(\Sigma^2) = \binom{n}{2}^{-1} \frac{1}{d}\left(1 - \frac{1}{d}\right).$$

Therefore, it is enough to show that if $n/\sqrt{d} \to \infty$, then the conditions of Theorem 2.2 are satisfied. To check the first condition, we calculate $\mathrm{tr}(\Sigma^4)$ and $\mathrm{tr}(\Sigma^2)$ as

$$\mathrm{tr}(\Sigma^4) = \frac{1}{d}\left(1 - \frac{1}{d}\right)\left\{\frac{1}{d^2}\left(1 - \frac{1}{d}\right) + \frac{1}{d^3}\right\} \quad \text{and} \quad \mathrm{tr}(\Sigma^2) = \frac{1}{d}\left(1 - \frac{1}{d}\right),$$

so that

$$\frac{\mathrm{tr}(\Sigma^4)}{\{\mathrm{tr}(\Sigma^2)\}^2} = \frac{1}{d} + \frac{1}{d(d-1)} \to 0 \quad \text{as } d \to \infty.$$

Next, we verify the second condition when $n/\sqrt{d} \to \infty$:

$$\frac{\mathbb{E}\left[\left\{h_I(\mathbf{X}_1, \mathbf{X}_2)\right\}^4\right] + n\mathbb{E}\left[\left\{h_I(\mathbf{X}_1, \mathbf{X}_2)\right\}^2\left\{h_I(\mathbf{X}_1, \mathbf{X}_3)\right\}^2\right]}{n^2\{\mathrm{tr}(\Sigma^2)\}^2} \to 0.$$

For the first part, we have

$$\begin{aligned}\mathbb{E}\left[\left\{h_I(\mathbf{X}_1, \mathbf{X}_2)\right\}^4\right] &= \mathbb{E}\left[\left\{(\mathbf{X}_1 - \pi_0)^\top(\mathbf{X}_2 - \pi_0)\right\}^4\right] \\ &= \frac{1}{d}\left(1 - \frac{1}{d}\right)\left\{\frac{1}{d^3} + \left(1 - \frac{1}{d}\right)^3\right\}.\end{aligned}$$

Therefore,

$$\frac{\mathbb{E}\left[\left\{h_I(\mathbf{X}_1, \mathbf{X}_2)\right\}^4\right]}{n^2\{\mathrm{tr}(\Sigma^2)\}^2} = \frac{\frac{1}{d^3}}{n^2\frac{1}{d}\left(1 - \frac{1}{d}\right)} + \frac{\left(1 - \frac{1}{d}\right)^3}{n^2\frac{1}{d}\left(1 - \frac{1}{d}\right)} \leq \frac{1}{n^2 d(d-1)} + \frac{d}{n^2} \;\to\; 0.$$

For the second part,

$$\begin{aligned}\mathbb{E}\left[\left\{h_I(\mathbf{X}_1, \mathbf{X}_2)\right\}^2\left\{h_I(\mathbf{X}_1, \mathbf{X}_3)\right\}^2\right] &= \mathbb{E}\left[\left\{(\mathbf{X}_1 - \pi_0)^\top(\mathbf{X}_2 - \pi_0)\right\}^2\left\{(\mathbf{X}_1 - \pi_0)^\top(\mathbf{X}_3 - \pi_0)\right\}^2\right] \\ &= \frac{1}{d^2}\left(1 - \frac{1}{d}\right)^2.\end{aligned}$$

This gives the second condition:

$$\frac{\mathbb{E}\left[\left\{h_I(\mathbf{X}_1, \mathbf{X}_2)\right\}^2 \left\{h_I(\mathbf{X}_1, \mathbf{X}_3)\right\}^2\right]}{n\{\mathrm{tr}(\Sigma^2)\}^2} = \frac{1}{n} \to 0.$$

Hence, the proof is complete.

A.1.7 Proof of Theorem 2.3

Note that the explicit formula for $\mathrm{Var}(U_A)$ is established in Lemma A.1.1. Recall the decomposition $U_A = U_{\mathrm{quad}} + U_{\mathrm{linear}}$ given in the main text. Then under $(S.1)$, we have

$$\frac{U_A - ||A^{1/2}(\pi - \pi_0)||_2^2}{\sqrt{\mathrm{Var}(U_A)}} = \frac{U_{\mathrm{linear}} - ||A^{1/2}(\pi - \pi_0)||_2^2}{\sqrt{\mathrm{Var}(U_{\mathrm{linear}})}} + o_P(1),$$

and the asymptotic normality follows by the usual central limit theorem. On the other hand, under (S.2), we have

$$\frac{U_A - ||A^{1/2}(\pi - \pi_0)||_2^2}{\sqrt{\mathrm{Var}(U_A)}} = \frac{U_{\mathrm{quad}}}{\sqrt{\mathrm{Var}(U_{\mathrm{quad}})}} + o_P(1).$$

Then we follow the similar steps in the proof of Theorem 2.2 to get the normality of U_{quad}. Hence the proof is complete.

A.1.8 Proof of Theorem 2.4

We proceed along the lines of the proof of Theorem 2 in Balakrishnan and Wasserman (2019). Note that the expectation and variance of U_w (Lemma A.1.1) are given by

$$\mathbb{E}\left[U_w\right] = ||A_w^{1/2}(\pi - \pi_0)||_2^2$$

$$\mathrm{Var}\left[U_w\right] = \binom{n}{2}^{-1} \Big\{\mathrm{tr}\{(A_w\Sigma)^2\} + 2(n-1)(\pi - \pi_0)^\top A_w \Sigma A_w (\pi - \pi_0)\Big\}.$$

Let $\mathbb{E}_0[\cdot]$, $\mathbb{E}_1[\cdot]$ be the expected value under the null and the alternative, respectively, and similarly denote $\mathrm{Var}_0[\cdot]$, $\mathrm{Var}_1[\cdot]$. By Chebyshev's inequality, under the null, we can see

$$\mathbb{P}_{H_0}\left(U_w \geq t_\alpha\right) \leq \frac{\mathrm{Var}_0\left[U_w\right]}{t_\alpha^2} = \alpha,$$

and $t_\alpha = \sqrt{\alpha^{-1}\mathrm{Var}_0\left[U_w\right]}$. This shows that $\phi(U_w)$ has size at most α.

For the type II error bound, assume the following two conditions are true:

$$(i)\ \ t_\alpha \le \frac{\mathbb{E}_1[U_w]}{2} \quad \text{and} \quad (ii)\ \ \sqrt{\frac{\mathrm{Var}_1[U_w]}{\zeta}} \le \frac{\mathbb{E}_1[U_w]}{2}.$$

Then, we can observe that

$$\begin{aligned}
\mathbb{P}_{H_1}(\phi(U_w)=0) &= \mathbb{P}_{H_1}(U_w < t_\alpha) \\
&\le \mathbb{P}_{H_1}\left(U_w < \frac{\mathbb{E}_1[U_w]}{2}\right) && \text{by } (i) \\
&= \mathbb{P}_{H_1}\left(U_w < \mathbb{E}_1[U_w] - \frac{\mathbb{E}_1[U_w]}{2}\right) \\
&\le \mathbb{P}_{H_1}\left(U_w < \mathbb{E}_1[U_w] - \sqrt{\frac{\mathrm{Var}_1[U_w]}{\zeta}}\right) && \text{by } (ii) \\
&\le \zeta,
\end{aligned}$$

where the last inequality follows by Chebyshev's inequality. Therefore, the proof can be done by showing (i) and (ii).

We begin with proving the first part (i). After some calculations, we can see

$$\mathrm{tr}\{(A_w\Sigma)^2\} = \sum_{j=1}^{d}\frac{\pi_j^2}{w_j^2} - 2\sum_{j=1}^{d}\frac{\pi_j^3}{w_j^2} + \left(\sum_{j=1}^{d}\frac{\pi_j^2}{w_j}\right)^2. \tag{A.12}$$

Therefore, under the null, the variance of U_w can be expanded to

$$\mathrm{Var}_0[U_w] = \binom{n}{2}^{-1}\left\{\sum_{j=1}^{d}\frac{\pi_{0,j}^2}{w_j^2} - 2\sum_{j=1}^{d}\frac{\pi_{0,j}^3}{w_j^2} + \left(\sum_{j=1}^{d}\frac{\pi_{0,j}^2}{w_j}\right)^2\right\}.$$

By Cauchy-Schwarz inequality, note that

$$\left(\sum_{j=1}^{d}\frac{\pi_j^2}{w_j}\right)^2 = \left(\sum_{j=1}^{d}\frac{\pi_j^{3/2}}{w_j}\pi_j^{1/2}\right)^2 \le \sum_{j=1}^{d}\frac{\pi_j^3}{w_j^2}\sum_{j=1}^{d}\pi_j = \sum_{j=1}^{d}\frac{\pi_j^3}{w_j^2}, \tag{A.13}$$

which implies

$$\mathrm{Var}_0[U_w] \le \binom{n}{2}^{-1}\sum_{j=1}^{d}\frac{\pi_{0,j}^2}{w_j^2}. \tag{A.14}$$

Using (A.14), the critical value is upper bounded by

$$t_\alpha = \sqrt{\alpha^{-1}\mathrm{Var}_0\left[U_w\right]} \;\leq \frac{2}{n}\sqrt{\frac{1}{\alpha}\sum_{j=1}^{d}\frac{\pi_{0,j}^2}{w_j^2}}$$

Note that, from the comparable condition, there exist $C_1, C_2 > 0$ and $\gamma \in (0,1)$ such that

$$C_1\{\gamma\pi_{0,i} + (1-\gamma)1/d\} \leq w_i \leq C_2\{\gamma\pi_{0,i} + (1-\gamma)1/d\} \quad \text{for all } i = 1, \ldots, d. \tag{A.15}$$

Consequently, the critical value is further upper bounded by

$$\begin{aligned} t_\alpha \leq \; & \frac{2}{n}\sqrt{\frac{1}{\alpha}\sum_{j=1}^{d}\frac{\pi_{0,j}^2}{w_j^2}} \;\leq\; \frac{2}{n}\sqrt{\frac{1}{\alpha}\sum_{j=1}^{d}\left(\frac{\pi_{0,j}}{C_1\{\gamma\pi_{0,i} + (1-\gamma)1/d\}}\right)^2} \\ \leq \; & \frac{2}{C_1\gamma n}\sqrt{\frac{d}{\alpha}}, \end{aligned} \tag{A.16}$$

where the last inequality is due to $1/\{C_1\gamma\pi_{0,i} + C_1(1-\gamma)1/d\} \leq 1/\{C_1\gamma\pi_{0,i}\}$. On the other hand, Cauchy-Schwarz inequality together with the comparable condition presents

$$\mathbb{E}_1\left[U_w\right] = \sum_{i=1}^{d}\frac{(\pi_i - \pi_{0,i})^2}{w_i} \;\geq\; \frac{||\pi - \pi_0||_1^2}{\sum_{i=1}^{d} w_i} \;\geq\; \frac{\epsilon_n^2}{C_2}. \tag{A.17}$$

Therefore, the first condition (i) is satisfied if

$$\epsilon_n^2 \;\geq\; \frac{4C_2}{C_1\gamma n}\sqrt{\frac{d}{\alpha}}.$$

This is the case from the assumption in (2.15).

Next, we prove the condition (ii). First, observe that

$$\mathrm{Var}_1\left[U_w\right] = \frac{1}{\binom{n}{2}}\Big\{\mathrm{tr}\{(A_w\Sigma)^2\} + 2(n-1)(\pi - \pi_0)^\top A_w\Sigma A_w(\pi - \pi_0)\Big\}.$$

By using the result in (A.12) and (A.13), the first trace term is bounded by

$$\frac{1}{\binom{n}{2}}\mathrm{tr}\{(A_w\Sigma)^2\} \;\leq\; \frac{4}{n^2}\sum_{j=1}^{d}\frac{\pi_j^2}{w_j^2},$$

for $n \geq 2$. On the other hand, the second term is bounded by

$$\begin{aligned}\frac{4}{n}(\pi-\pi_0)^\top A_w \Sigma A_w(\pi-\pi_0) &= \frac{4}{n}(\pi-\pi_0)^\top A_w \left(\mathrm{diag}\{\pi\}-\pi\pi^\top\right) A_w(\pi-\pi_0) \\ &\leq \frac{4}{n}(\pi-\pi_0)^\top A_w \mathrm{diag}\{\pi\} A_w(\pi-\pi_0) = \frac{4}{n}\sum_{j=1}^d \frac{\Delta_j^2\pi_j}{w_j^2},\end{aligned}$$

where $\Delta_i = \pi_{0,i} - \pi_i$. Therefore, we have

$$\begin{aligned}\mathrm{Var}_1[U_w] &\leq \frac{4}{n^2}\sum_{j=1}^d \frac{\pi_j^2}{w_j^2} + \frac{4}{n}\sum_{j=1}^d \frac{\Delta_j^2\pi_j}{w_j^2} \\ &= \frac{4}{n^2}\sum_{j=1}^d \frac{\pi_{0,j}^2+\Delta_j^2-2\pi_{0,j}\Delta_j}{w_j^2} + \frac{4}{n}\sum_{j=1}^d \frac{\Delta_j^2\pi_{0,j}-\Delta_j^3}{w_j^2} \\ &\leq \underbrace{\frac{8}{n^2}\sum_{j=1}^d \frac{\pi_{0,j}^2}{w_j^2}}_{U_1} + \underbrace{\frac{8}{n^2}\sum_{j=1}^d \frac{\Delta_j^2}{w_j^2}}_{U_2} + \underbrace{\frac{8}{n}\sum_{j=1}^d \frac{\Delta_j^2\pi_{0,j}}{w_j^2}}_{U_3} + \underbrace{\frac{8}{n}\sum_{j=1}^d \frac{|\Delta_j|^3}{w_j^2}}_{U_4}.\end{aligned}$$

To finish the proof, we need to verify

$$\sum_{i=1}^4 \frac{2\sqrt{U_i/\zeta}}{\mathbb{E}_1[U_w]} \leq 1.$$

Indeed, this is the case by modifying the result in Balakrishnan and Wasserman (2019) with a different constant factor. To show the details, using (A.17), the first term is upper bounded by

$$\frac{2\sqrt{U_1/\zeta}}{\mathbb{E}_1[U_w]} \leq \frac{4\sqrt{2}C_2}{\sqrt{\zeta}n\epsilon_n^2}\sqrt{\sum_{j=1}^d \frac{\pi_{0,j}^2}{w_j^2}} \leq \frac{4\sqrt{2}C_2}{\sqrt{\zeta}C_1\gamma}\frac{\sqrt{d}}{n\epsilon_n^2} \leq \frac{1}{4}.$$

For the second term, note that

$$\begin{aligned}U_2 &= \frac{8}{n^2}\sum_{j=1}^d \frac{\Delta_j^2}{w_j^2} \leq \frac{8}{n^2}\sum_{j=1}^d \frac{\Delta_j^2}{C_1^2\{\gamma\pi_{0,j}+(1-\gamma)1/d\}^2} = \frac{8d^2}{n^2C_1^2(1-\gamma)^2}\sum_{j=1}^d \frac{\Delta_j^2}{\{d\pi_{0,j}\gamma/(1-\gamma)+1\}^2} \\ &\leq \frac{8d^2}{n^2C_1^2(1-\gamma)^2}\sum_{j=1}^d \frac{\Delta_j^2}{\{d\pi_{0,j}\gamma/(1-\gamma)+1\}} = \frac{8d}{n^2C_1^2(1-\gamma)}\underbrace{\sum_{j=1}^d \frac{\Delta_j^2}{\gamma\pi_{0,j}+(1-\gamma)1/d}}_{\overset{let}{=}\rho_n}.\end{aligned}$$

The expected value is lower bounded in terms of ρ_n by

$$\mathbb{E}_1\left[U_w\right] = \sum_{j=1}^{d} \frac{\Delta_j^2}{w_j} \geq \frac{\rho_n}{C_2},$$

and similarly to (A.17), it is seen that $\rho_n \geq \epsilon_n^2$. Using these results,

$$\frac{2\sqrt{U_2/\zeta}}{\mathbb{E}_1\left[U_w\right]} \leq \frac{4C_2\sqrt{2d}}{C_1\sqrt{\zeta(1-\gamma)n\epsilon_n}} \leq \frac{1}{4}. \tag{A.18}$$

For the third term, note that

$$\frac{\pi_{0,j}}{w_j} \leq \frac{\pi_{0,j}}{C_1\{\gamma\pi_{0,j} + (1-\gamma)1/d\}} \leq \frac{1}{C_1\gamma}.$$

Using this inequality,

$$U_3 = \frac{8}{n}\sum_{j=1}^{d} \frac{\Delta_j^2\pi_{0,j}}{w_j^2} \leq \frac{8}{C_1\gamma n}\sum_{j=1}^{d} \frac{\Delta_j^2}{w_j} = \frac{8}{C_1\gamma n}\mathbb{E}_1\left[U_w\right].$$

As a result,

$$\frac{2\sqrt{U_3/\zeta}}{\mathbb{E}_1\left[U_w\right]} \leq \frac{2\sqrt{2C_2}}{\sqrt{C_1\gamma\zeta n}\epsilon_n} \leq \frac{1}{4}.$$

To control the last term, the monotonicity of the ℓ_p norms and the comparable condition present

$$U_4 = \frac{8}{n}\sum_{j=1}^{d} \frac{|\Delta_j|^3}{w_j^2} \leq \frac{8}{n}\left(\sum_{j=1}^{d} \frac{\Delta_j^2}{w_j^{4/3}}\right)^{3/2} \leq \frac{8d^{1/2}}{nC_1^2(1-\gamma)^{1/2}}\left(\sum_{j=1}^{d} \frac{\Delta_j^2}{\gamma\pi_{0,j} + (1-\gamma)1/d}\right)^{3/2}.$$

Based on the result,

$$\frac{2\sqrt{U_4/\zeta}}{\mathbb{E}_1\left[U_w\right]} \leq \frac{4\sqrt{2}C_2d^{1/4}}{C_1(1-\gamma)^{1/4}\sqrt{\zeta n}\rho_n^{1/4}} \leq \frac{4\sqrt{2}C_2d^{1/4}}{C_1(1-\gamma)^{1/4}\sqrt{\zeta n}\epsilon_n^{1/2}} \leq \frac{1}{4}.$$

This completes the proof.

A.2 Asymptotics under Poissonization

It is a common assumption in the literature on multinomial testing (Diakonikolas et al., 2015; Valiant and Valiant, 2017; Balakrishnan and Wasserman, 2019) that the sample size n' has a Poisson distribution with parameter n. In this case, the number of occurrences in the i^{th} category has an independent Poisson($n\pi_i$).

This approach, so called *Poissonization*, makes the analysis simple and straightforward. In this section, we study the asymptotic behavior of some variants of chi-square statistic under Poissonization.

Let n' be a random sample from Poisson(n) and draw n' independent samples from a multinomial distribution with parameters $\pi = (\pi_1, \ldots, \pi_d)$. Denote the number of observations in the j^{th} category by $Y_j = \sum_{i=1}^{n'} I(X_{i,j} = 1)$, which has an independent Poisson$(n\pi_j)$. Given positive weights $(w_1, \ldots, w_d)$, a weighted Poissonized chi-square statistic is defined by

$$T_w = \sum_{j=1}^{d} \frac{(Y_j - n\pi_{0,j})^2 - Y_j}{w_j}. \tag{A.19}$$

For instance, if w_j is given as $w_j = \pi_{0,j}^{2/3}$, then T_w corresponds to the test statistic by Valiant and Valiant (2017), and if $w_j = \max\{1/d, \pi_{0,j}\}$, it corresponds to the test statistic by Balakrishnan and Wasserman (2019). Since $Y_1, \ldots, Y_d$ are independent, it is rather straightforward to have the asymptotic normality of T_w under the null as $n, d \to \infty$. First, note, under the null, that the expected value of T_w is zero, and the variance of T_w can be obtained by

$$\mathrm{Var}(T_w) = \sum_{j=1}^{d} \frac{2(n\pi_{0,j})^2}{w_j^2}. \tag{A.20}$$

The next theorem provides a sufficient condition that leads to the asymptotic normality of T_w in the high-dimensional regime.

Theorem A.2 (Asymptotic normality of T_w under the null)**.** *Let us denote the variance of* T_w *by* $\sigma_{n,d}^2 = \mathrm{Var}(T_w)$ *in (A.20). If*

$$\lim_{n,d\to\infty} \frac{1}{\sigma_{n,d}^4} \sum_{j=1}^{d} \frac{60(n\pi_{0,j})^4 + 144(n\pi_{0,j})^3 + 8(n\pi_{0,j})^2}{w_j^4} = 0, \tag{A.21}$$

then, under the null,

$$\sigma_{n,d}^{-1} T_w \xrightarrow{d} \mathcal{N}(0, 1). \tag{A.22}$$

Proof. Since Y_j has an independent Poisson$(n\pi_{0,j})$ under the null, a straightforward but involved calculations presents

$$\mathbb{E}\left[\left\{\frac{(Y_j - n\pi_{0,j})^2 - Y_j}{w_j}\right\}^4\right] = \frac{60(n\pi_{0,j})^4 + 144(n\pi_{0,j})^3 + 8(n\pi_{0,j})^2}{w_j^4}.$$

The proof is completed by the Lyapounovs condition with $\delta = 2$. □

Note that Theorem A.2 is a generalization of Lemma 6 of Balakrishnan and Wasserman (2019) where they assume $\pi_{0,j} = d^{-1}$ and $w_j = 1$ for all $j = 1, \ldots, d$. In the uniform null case, a sufficient condition for the normal approximation is $n/\sqrt{d} \to \infty$. The next theorem shows that if $n/\sqrt{d} \to c \in (0, \infty)$, T_w converges to a Poisson distribution under the uniform null, which is analogous to Corollary 2.1.1 for U_I.

Theorem A.3 (Poissonian asymptotic for T_w under the uniform null)**.** *Suppose the weights* $(w_1, \ldots, w_d)$ *given in (A.19) have the same value under the uniform null. Without loss of generality, let* $w_j = 2n/\sqrt{d}$ *for all* $j = 1, \ldots, d$*, and assume* $n/\sqrt{d} \to c \in (0, \infty)$*. Then we have, under the uniform null,*

$$T_w \xrightarrow{d} \frac{1}{c}\text{Poisson}\left(\frac{c^2}{2}\right) - \frac{c}{2}.$$

Proof. Note that T_w can be decomposed into

$$T_w = \underbrace{\frac{\sqrt{d}}{n}\sum_{j=1}^{d}\frac{Y_j(Y_j-1)}{2}}_{D_1} - \underbrace{\frac{1}{\sqrt{d}}\sum_{j=1}^{d}Y_j + \frac{n}{2\sqrt{d}}}_{D2}.$$

Since Y_j has an independent Poisson(n/d), we have $\text{Var}\,(D_2) = n/d \to 0$ under the given assumption, which in turn implies $D_2 \xrightarrow{p} -c/2$. Hence, it is sufficient to show $D_1 \xrightarrow{d} c^{-1}\text{Poisson}(c^2/2)$.

In order to show the Poisson limit, we use the Stein-Chen's method for Poisson approximations (e.g., Barbour et al., 1992). Let us denote

$$W = \sum_{j=1}^{d}\frac{Y_j(Y_j-1)}{2} = \sum_{j=1}^{d}\psi_j,$$

where $\psi_j = Y_j(Y_j-1)/2$. At a high-level, if $n/\sqrt{d} \to c$, then ψ_j behaves like an independent indicator random variable, and the result follows by the law of small numbers. To start with, let $\lambda = \mathbb{E}\,[T_w]$ and consider a function $f = f_{\lambda,A} : \mathbb{Z}^+ \to \mathbb{R}$ which is the solution of

$$\lambda f(j+1) - jf(j) = I\,(j \in A) - \text{Poi}_\lambda(A), \quad j \geq 0, \tag{A.23}$$

for all $A \in \mathbb{Z}^+$. From e.g., Lemma 1.1.1 of Barbour et al. (1992), the solution of (A.23) satisfies

$$\sup_j |f_{\lambda,A}(j)| \leq \min\left(1, \lambda^{-1/2}\right), \quad \sup_j |f_{\lambda,A}(j+1) - f_{\lambda,A}(j)| \leq \min\left(1, \lambda^{-1}\right). \tag{A.24}$$

Let $\mathcal{F}$ be a class of functions that satisfies (A.24), and Z be a Poisson random variable with λ. Then it is clear from (A.23) to see that

$$d_{TV}(W,Z) \leq \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]\right|. \tag{A.25}$$

For the rest of the proof, we will bound the right-side of (A.25). First, observe that

$$\begin{aligned}
\mathbb{E}\left[Wf(W)\right] &= \sum_{j=1}^{d} \mathbb{E}\left[\psi_j f(W)\right] \\
&= \sum_{j=1}^{d} \mathbb{E}\left[f(W)\big|\psi_j = 1\right]\mathbb{P}(\psi_j = 1) + \underbrace{\sum_{j=1}^{d}\sum_{k=2}^{\infty} \mathbb{E}\left[\psi_j f(W)\big|\psi_j = k\right]\mathbb{P}\left(\psi_j = k\right)}_{R_d}.
\end{aligned}$$

Since $\mathbb{P}(\psi_j = 1) = \mathbb{P}(Y_j = 2)$,

$$\begin{aligned}
\mathbb{E}\left[Wf(W)\right] &= \sum_{j=1}^{d} \mathbb{E}\left[f(W)\big|\psi_j = 1\right]\frac{(n/d)^2 e^{-n/d}}{2} + R_d \\
&= \sum_{j=1}^{d} \mathbb{E}\left[f(W_j+1)\big|\psi_j = 1\right]\frac{(n/d)^2 e^{-n/d}}{2} + R_d \\
&= \sum_{j=1}^{d} \mathbb{E}\left[f(W_j+1)\right]\frac{(n/d)^2 e^{-n/d}}{2} + R_d,
\end{aligned} \tag{A.26}$$

where $W_j = W - \psi_j$, and (A.26) follows by $W_j \perp\!\!\!\perp \psi_j$. Now, we have

$$\begin{aligned}
\left|\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]\right| &= \left|\left[\sum_{j=1}^{d}\frac{n^2}{2d^2}\mathbb{E}\Big\{f(W+1) - e^{-\frac{n}{d}} f(W_j+1)\Big\}\right] + R_d\right| \\
&= \left|\left[\sum_{j=1}^{d}\frac{n^2}{2d^2}\mathbb{E}\Big\{f(W+1) - f(W_j+1) + f(W_j+1) - e^{-\frac{n}{d}} f(W_j+1)\Big\}\right] + R_d\right| \\
&\leq \underbrace{\left|\left[\sum_{j=1}^{d}\frac{n^2}{2d^2}\mathbb{E}\Big\{f(W+1) - f(W_j+1)\Big\}\right]\right|}_{(i)} + \underbrace{\left|\left[\sum_{j=1}^{d}\frac{n^2}{2d^2}\mathbb{E}\Big\{f(W_j+1) - e^{-\frac{n}{d}} f(W_j+1)\Big\}\right]\right|}_{(ii)} + R_d.
\end{aligned}$$

For the part (i), by telescoping and $||\Delta f|| \leq 1$, we have

$$\left|\left[\sum_{j=1}^{d}\frac{n^2}{2d^2}\mathbb{E}\Big\{f(W+1) - f(W_j+1)\Big\}\right]\right| \;\leq\; \sum_{j=1}^{d}\frac{n^2}{2d^2}||\Delta f||\mathbb{E}\big|W - W_j\big|$$

$$\leq \frac{n^2}{2d}\mathbb{E}\left[\psi_1\right] = \frac{n^4}{4d^3}. \tag{A.27}$$

For the part (ii), by using the triangle inequality and $|f| \leq 1$, we have

$$\left|\left[\sum_{j=1}^{d} \frac{n^2}{2d^2}\mathbb{E}\Big\{f(W_j+1) - e^{-\frac{n}{d}} f(W_j+1)\Big\}\right]\right| \leq \frac{n^2}{2d}\left(1 - e^{-\frac{n}{d}}\right). \tag{A.28}$$

Lastly, we control the remainder term:

$$\begin{aligned}
R_d &= \sum_{j=1}^{d}\sum_{k=2}^{\infty} \mathbb{E}\left[\psi_j f(W) \big| \psi_j = k\right] \mathbb{P}\left(\psi_j = k\right) \\
&\leq \sum_{j=1}^{d}\sum_{k=2}^{\infty} k\mathbb{P}\left(\psi_j = k\right) = \sum_{j=1}^{d}\sum_{k=3}^{\infty} k\mathbb{P}\left(Y_j = \frac{1+\sqrt{1+8k}}{2}\right) \\
&= \sum_{j=1}^{d}\sum_{k=3}^{\infty} \frac{k(k-1)}{2}\mathbb{P}\left(Y_j = k\right) \leq \sum_{j=1}^{d}\sum_{k=3}^{\infty} k^2\mathbb{P}\left(Y_j = k\right) \\
&= \sum_{j=1}^{d}\Big\{\mathbb{E}\left[Y_j^2\right] - \mathbb{P}(Y_j = 1) - 4\mathbb{P}(Y_j = 2)\Big\} = \sum_{j=1}^{d}\Big\{\frac{n}{d} + \frac{n^2}{d^2} - \frac{n}{d}e^{-\frac{n}{d}} - 2\frac{n^2}{d^2}e^{-\frac{n}{d}}\Big\} \\
&= n\left(1 - e^{-\frac{n}{d}}\right) + \frac{n^2}{d}\left(1 - 2e^{-\frac{n}{d}}\right).
\end{aligned}$$

Combining (A.27), (A.28) and (A.29),

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[\lambda f(W+1) - Wf(W)\right]\right| &\leq \frac{n^4}{4d^3} + \frac{n^2}{2d}\left(1 - e^{-\frac{n}{d}}\right) + n\left(1 - e^{-\frac{n}{d}}\right) + \frac{n^2}{d}\left(1 - 2e^{-\frac{n}{d}}\right) \\
&\to 0 \quad \text{as} \quad n, d \to \infty \quad \text{and} \quad n/\sqrt{d} \to c.
\end{aligned}$$

It follows that $d_{TV}(W, Z) \to 0$, and consequently, $W \to \text{Poisson}(c^2/2)$. □

Let us consider a diagonal weight matrix $A_w = \text{diag}\{w_1^{-1}, \ldots, w_d^{-1}\}$. The next lemma provides the mean and variance of T_w for a general case.

Lemma A.3.1. *Let us denote* $\text{diag}(\pi)$ *by* Γ*, and* $\Delta = \pi - \pi_0$*. Then*

$$\mathbb{E}\left[T_w\right] = n^2||A_w^{1/2}\Delta||_2^2 \quad \textit{and} \quad \text{Var}(T_w) = 2n^2\text{tr}\big\{(A_w\Gamma)^2\big\} + 4n^3\Delta^\top A_w\Gamma A_w\Delta. \tag{A.29}$$

We briefly investigate the power of T_w under the Gaussian regime. As in Section 2.3.2, we begin by providing the two scenarios at the alternative hypothesis, and derive the limiting distribution of T_w under the considered scenarios. Analogous to (S.1) and (S.2), we define

(S′.1) (Strong Signal-to-Noise) $n^{-1}\mathrm{tr}((A_w\Gamma)^2) = o\big((\pi-\pi_0)^\top A_w\Gamma A_w(\pi-\pi_0)\big)$

(S′.2) (Weak Signal-to-Noise) $(\pi-\pi_0)^\top A_w\Gamma A_w(\pi-\pi_0) = o\big(n^{-1}\mathrm{tr}((A_w\Gamma)^2)\big)$

where $\Gamma = \mathrm{diag}\,(\pi)$. We then present the following theorem on the asymptotic distribution of T_w under the given alternatives.

Theorem A.4 (Asymptotic normality of T_w under the alternative)**.** *Assume either i)* (S′.1) *and* $(\pi-\pi_0)^\top A_w\Gamma A_w(\pi-\pi_0) < \infty$ *or ii)* (S′.2) *and*

$$\lim_{n,d\to\infty} \frac{1}{\varsigma_{n,d}^4} \sum_{j=1}^{d} \frac{60(n\pi_j)^4 + 144(n\pi_j)^3 + 8(n\pi_j)^2}{w_j^4} = 0, \tag{A.30}$$

where $\varsigma_{n,d}^2 = 2n^2\sum_{j=1}^{d}(\pi_j/w_j)^2$*. Then, under the considered alternatives, we have*

$$\frac{T_w - n^2||A_w^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{\mathrm{Var}(T_w)}} \xrightarrow{d} \mathcal{N}(0,1). \tag{A.31}$$

Proof. Note that the numerator of the standardized statistic is

$$\begin{aligned} T_w - n^2||A_w^{1/2}(\pi-\pi_0)||_2^2 &= \sum_{j=1}^{d} \frac{(Y_j - n\pi_j)^2 - Y_j}{w_j} + \sum_{j=1}^{d} \frac{2n(\pi_j-\pi_{0,j})(Y_j - n\pi_j)}{w_j} \\ &= T_{w,\mathrm{quad}} + T_{w,\mathrm{linear}}, \end{aligned}$$

and the variance of each term is given by

$$\begin{aligned} &\mathrm{Var}\,(T_{w,\mathrm{quad}}) = 2n^2\mathrm{tr}((A_w\Gamma)^2) = \varsigma_{n,d}^2 \\ &\mathrm{Var}\,(T_{w,\mathrm{linear}}) = 4n^3(\pi-\pi_0)^\top A_w\Gamma A_w(\pi-\pi_0) \\ &\mathrm{Cov}\,(T_{w,\mathrm{quad}}, T_{w,\mathrm{linear}}) = 0. \end{aligned}$$

Therefore, under (S′.1),

$$\frac{T_w - n^2||A_w^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{\mathrm{Var}(T_w)}} = \frac{T_{w,\mathrm{linear}}}{\sqrt{\mathrm{Var}(T_{w,\mathrm{linear}})}} + o_P(1) \xrightarrow{d} \mathcal{N}(0,1),$$

which is followed by the usual central limit theorem.

Under (S′.2) and the assumption (A.30),

$$\frac{T_w - n^2||A_w^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{\mathrm{Var}(T_w)}} = \frac{T_{w,\mathrm{quad}}}{\sqrt{\mathrm{Var}(T_{w,\mathrm{quad}})}} + o_P(1) \xrightarrow{d} \mathcal{N}(0,1),$$

which is followed by Theorem A.2. This finishes the proof. □

As in Section 2.3.2, we further assume

(S′.3) $n^{-1}\mathrm{tr}((A_w\Gamma_0)^2) = o\big((\pi-\pi_0)^\top A_w\Gamma A_w(\pi-\pi_0)\big)$,

to simplify the power function. Then under (S′.1) and (S′.3), the power of T_w is approximated by

$$\beta'_{n,d}(\pi_0,\pi_1,A_w) = \Phi\left(\frac{\sqrt{n}||A_w^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{4(\pi-\pi_0)^\top A_w\Gamma A_w(\pi-\pi_0)}}\right) + o(1).$$

Whereas, under (S′.2), we have

$$\beta'_{n,d}(\pi_0,\pi_1,A_w) = \Phi\left(-\frac{\sqrt{\mathrm{tr}\{(A_w\Gamma_0)^2\}}}{\sqrt{\mathrm{tr}\{(A_w\Gamma)^2\}}}z_\alpha + \frac{n||A_w^{1/2}(\pi-\pi_0)||_2^2}{\sqrt{2\mathrm{tr}\{(A_w\Gamma)^2\}}}\right) + o(1).$$

Appendix B

Appendix for Chapter 3

B.1 Proofs

B.1.1 Proof of Theorem 3.1

We start by simplifying $\widehat{m}_{\mathrm{LDA}}(x)$ as

$$\begin{aligned}
& \widehat{m}_{\mathrm{LDA}}(X_i) \\
= & \frac{\pi_1 \exp\big\{-\frac{1}{2}(X_i-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_1)\big\}}{\pi_1 \exp\big\{-\frac{1}{2}(X_i-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_1)\big\} + \pi_0 \exp\big\{-\frac{1}{2}(X_i-\widehat{\mu}_0)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_0)\big\}} \\
= & \frac{\pi_1}{\pi_1 + \pi_0 \exp\big\{-\frac{1}{2}(X_i-\widehat{\mu}_0)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_0) + \frac{1}{2}(X_i-\widehat{\mu}_1)^\top \mathcal{S}^{-1}(X_i-\widehat{\mu}_1)\big\}} \\
= & \frac{\pi_1}{\pi_1 + \pi_0 \exp\Big\{\left(X_i - (\widehat{\mu}_0+\widehat{\mu}_1)/2\right)^\top \mathcal{S}^{-1}\left(\widehat{\mu}_0-\widehat{\mu}_1\right)\Big\}}
\end{aligned}$$

and write

$$W_i = \left(X_i - (\widehat{\mu}_0+\widehat{\mu}_1)/2\right)^\top \mathcal{S}^{-1}\left(\widehat{\mu}_0-\widehat{\mu}_1\right).$$

For some $a \in (0,1)$, Taylor expansion of $f(x) = a/\{a+(1-a)e^x\}$ at $x=0$ provides

$$\left|\left\{\widehat{m}_{\mathrm{LDA}}(X_i)-\pi_1\right\}^2 - \pi_0^2\pi_1^2 W_i^2\right| \le C|W_i|^3,$$

where C is a universal constant. This implies that

$$\left| \sum_{i=1}^{n} \left\{ \widehat{m}_{\mathrm{LDA}}(X_i) - \pi_1 \right\}^2 - \pi_0^2 \pi_1^2 \sum_{i=1}^{n} W_i^2 \right| \;\leq\; C \sum_{i=1}^{n} |W_i|^3.$$

Now based on $|x+y|^3 \leq 4|x|^3 + 4|y|^3$ and Cauchy-Schwarz inequality, it can be seen that

$$\sum_{i=1}^{n} |W_i|^3 \leq 4n \left| ((\widehat{\mu}_0 + \widehat{\mu}_1)/2)^\top \mathcal{S}^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1) \right|^3 + 4 \sum_{i=1}^{n} \left| X_i^\top \mathcal{S}^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1) \right|^3 = o_P(1).$$

As a result, $n\widehat{\mathcal{T}}_{\mathrm{LDA}}$ can be approximated by

$$n\widehat{\mathcal{T}}_{\mathrm{LDA}} = \sum_{i=1}^{n} \left\{ \widehat{m}_{\mathrm{LDA}}(X_i) - \pi_1 \right\}^2 = \pi_0^2 \pi_1^2 \sum_{i=1}^{n} W_i^2 + o_P(1). \tag{B.1}$$

Let us denote $\delta_n = \mathcal{S}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1)$ and $\Delta_n = (\widehat{\mu}_0 + \widehat{\mu}_1)/2$, and recall $\mathcal{S} = n^{-1} \sum_{i=1}^{n} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$ where $\widehat{\mu} = n^{-1} \sum_{i=1}^{n} X_i$. Then we observe that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} W_i^2 \;&=\; \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_n^\top X_i - \delta_n^\top \Delta_n \right\}^2 \\
&=\; \delta_n^\top \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_i - \Delta_n)(X_i - \Delta_n)^\top \right\} \delta_n \\
&=\; \delta_n^\top \mathcal{S} \delta_n + \delta_n^\top (\widehat{\mu} - \Delta_n)(\widehat{\mu} - \Delta_n)^\top \delta_n \\
&=\; (\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1) + R_n,
\end{aligned}$$

where $R_n = \delta_n^\top (\widehat{\mu} - \Delta_n)(\widehat{\mu} - \Delta_n)^\top \delta_n$. Hence, we have

$$n\widehat{\mathcal{T}}_{\mathrm{LDA}} = n\pi_0^2 \pi_1^2 \left\{ (\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1) + R_n \right\} + o_P(1).$$

We also note that the residual term is negligible under the null, i.e. $n\pi_0^2 \pi_1^2 R_n = o_P(1)$, which results in

$$\begin{aligned}
n\pi_0^{-1} \pi_1^{-1} \widehat{\mathcal{T}}_{\mathrm{LDA}} &= \frac{n_0 n_1}{n_0 + n_1} (\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1) + o_P(1) \\
&= T^2_{\mathrm{Hotelling}} + o_P(1).
\end{aligned}$$

The rest of the proof follows by the limiting property of Hotelling's T^2.

B.1.2 Proof of Theorem 3.2

Proof. First note that the likelihood ratio for testing (3.8) is given by

$$\mathcal{L}_n = \sum_{i=1}^{n_1} \log \frac{f_{\mu_0+h/\sqrt{n}}(X_{i,1})}{f_{\mu_0}(X_{i,1})}.$$

Since $\{\mathbb{P}_\mu, \mu \in \Omega\}$ is q.m.d. at μ_0, Theorem 12.2.3 of Lehmann and Romano (2006) under $n_1/(n_0+n_1) \to \pi_1$ yields that

$$\mathcal{L}_n \xrightarrow{d} N\left(-\frac{\pi_1}{2}\langle h, I(\mu_0)h\rangle,\ \pi_1\langle h, I(\mu_0)h\rangle\right),$$

where $I(\mu)$ is the Fisher information matrix. This implies by Corollary 12.3.1 of Lehmann and Romano (2006) that the joint distribution of $X_{1,0}$ and $X_{1,1}$ under the null and the alternative are mutually contiguous. Since contiguity implies

$$n\pi_0^{-1}\pi_1^{-1}\widehat{\mathcal{T}}_{\mathrm{LDA}} = \frac{n_0 n_1}{n_0+n_1}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \mathcal{S}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) + o_P(1),$$

under $H_{1,n}$, the result follows by the limiting distribution of Hotelling's T^2 statistic. □

B.1.3 Proof of Theorem 3.3

Proof. The exact type I error control of the permutation test is well-known (see e.g. Chapter 15 of Lehmann and Romano, 2006). Strictly speaking, the considered test is not the usual permutation test since the only first half of labels are permuted to decide a critical value. However, it also controls the type I error under H_0 due to Theorem 15.2.1 of Lehmann and Romano (2006). Indeed, this result holds regardless of i.i.d. sampling or separate sampling. Hence we focus on the type II error control.

• Type II error control (i.i.d. sampling)

We start with the case of i.i.d. sampling. Based on the inequality $(x-y)^2 \le 2(x-z)^2 + 2(z-y)^2$, we lower bound the test statistic as

$$\begin{aligned}
\widehat{\mathcal{T}}'_{global} &= \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - \widehat{\pi}_1)^2 \\
&\ge \frac{1}{2n}\sum_{i=n+1}^{2n}(m(X_i) - \widehat{\pi}_1)^2 - \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - m(X_i))^2 \\
&\ge \frac{1}{4n}\sum_{i=n+1}^{2n}(m(X_i) - \pi_1)^2 - \frac{1}{2}(\pi_1 - \widehat{\pi}_1)^2 - \frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - m(X_i))^2. \qquad \text{(B.2)}
\end{aligned}$$

Define the events $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4$ such that

$$\mathcal{A}_1 = \Big\{(\pi_1 - \widehat{\pi}_1)^2 < C_2\delta_n\Big\},$$

$$\mathcal{A}_2 = \Big\{\frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}(X_i) - m(X_i))^2 < C_3\delta_n\Big\},$$

$$\mathcal{A}_3 = \Big\{\Big|\frac{1}{n}\sum_{i=n+1}^{2n}(m(X_i) - \pi_1)^2 - \mathbb{E}\left[(m(X_i) - \pi_1)^2\right]\Big| < \frac{1}{2}\mathbb{E}\left[(m(X) - \pi_1)^2\right]\Big\},$$

$$\mathcal{A}_4 = \Big\{t_\alpha < C'_{0,\alpha}\delta_n\Big\}.$$

Using Markov's inequality, we have

$$\mathbb{P}(\mathcal{A}_1^c) \leq \frac{\pi_1(1-\pi_1)}{C_2 n\delta_n},$$

$$\mathbb{P}(\mathcal{A}_2^c) \leq \frac{1}{C_3\delta_n}\mathbb{E}\left[\int_S (\widehat{m}(x) - m(x))^2 dP_X(x)\right] \leq \frac{C_0}{C_3},$$

by the condition in (3.9). For the third event, denote $\Delta_n = \mathbb{E}\left[(m(X) - \pi_1)^2\right]$ and use Chebyshev's inequality to have

$$\begin{aligned}
\mathbb{P}(\mathcal{A}_3^c) &\leq \frac{4}{n\Delta_n^2}\mathrm{Var}\left[(m(X) - \pi_1)^2\right] \\
&\leq \frac{4}{n\Delta_n^2}\mathbb{E}\left[(m(X) - \pi_1)^4\right] \\
&\leq \frac{4}{n\Delta_n^2}\mathbb{E}\left[(m(X) - \pi_1)^2\right] \quad \text{since } |m(X) - \pi_1| \leq 1 \\
&\leq \frac{4}{C_1 n\delta_n},
\end{aligned}$$

where the last inequality uses the assumption that $\Delta_n \geq C_1\delta_n$. Furthermore, under the assumption on the permutation critical value, $\mathbb{P}(\mathcal{A}_4^c) \leq \beta/2$. Hence, we obtain

$$\mathbb{P}\left((\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4)^c\right) \leq \sum_{i=1}^{4}\mathbb{P}(\mathcal{A}_i^c) < \beta,$$

by choosing sufficiently large $C_1, C_2, C_3 > 0$ with the assumption that $\delta_n \geq n^{-1}$. Using (B.2), the type II error of the regression test is bounded by

$$\mathbb{P}(\widehat{\mathcal{T}}'_{global} \leq t_\alpha)$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(\frac{1}{4n} \sum_{i=n+1}^{2n} (m(X_i) - \pi_1)^2 - \frac{1}{2}(\pi_1 - \widehat{\pi}_1)^2 - \frac{1}{n} \sum_{i=n+1}^{2n} (\widehat{m}(X_i) - m(X_i))^2 \leq t_\alpha \right) \\
&\leq \mathbb{P}\left(\left\{ \frac{1}{4n} \sum_{i=n+1}^{2n} (m(X_i) - \pi_1)^2 - \frac{1}{2}(\pi_1 - \widehat{\pi}_1)^2 - \frac{1}{n} \sum_{i=n+1}^{2n} (\widehat{m}(X_i) - m(X_i))^2 \leq t_\alpha \right\} \right. \\
&\qquad \left. \bigcap \left\{ \bigcap_{j=1}^{4} \mathcal{A}_j \right\} \right) + \mathbb{P}\left((\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3 \cap \mathcal{A}_4)^c \right) \\
&\leq \mathbb{P}\left(\Delta_n < C_4 \delta_n \right) + \beta,
\end{aligned}
$$

where C_4 can be chosen by $C_4 = 2C'_{0,\alpha} + C_2 + 2C_3$. Now by choosing $C_1 > C_4$ for sufficiently large n, the type II error can be bounded by β. Hence the result follows.

• **Type II error control (Separate Sampling)**

The proof for separate sampling is almost the same as before except few details. First, we do not need to define $\mathcal{A}_1$ since π_1 is known. In terms of $\mathcal{A}_2$, apply Markov's inequality to obtain

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{A}_2^c\right) &\leq \frac{1}{C_3 \delta_n} \left\{ \frac{n_0}{n} \mathbb{E}\left[\int_S (\widehat{m}(x) - m(x))^2 dP_0(x) \right] + \frac{n_1}{n} \mathbb{E}\left[\int_S (\widehat{m}(x) - m(x))^2 dP_1(x) \right] \right\} \\
&= \frac{C_0}{C_3} \mathbb{E}\left[\int_S (\widehat{m}(x) - m(x))^2 dP_X(x) \right] \leq \frac{C_0}{C_3},
\end{aligned}
$$

where the last line uses the fact that $\frac{n_0}{n} P_0 + \frac{n_1}{n} P_1 = P_X$. Similarly, for the event $\mathcal{A}_3$, we have by Chebyshev's inequality that

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{A}_3^c\right) &\leq \frac{4}{\Delta_n^2} \frac{1}{n^2} \sum_{i=n+1}^{2n} \mathrm{Var}\left[(m(X_i) - \pi_1)^2 \right] \\
&\leq \frac{4}{\Delta_n^2} \frac{1}{n^2} \sum_{i=n+1}^{2n} \mathbb{E}\left[(m(X_i) - \pi_1)^2 \right] = \frac{4}{n \Delta_n^2} \mathbb{E}\left[(m(X) - \pi_1)^2 \right] \\
&\leq \frac{4}{C_1 n \delta_n}.
\end{aligned}
$$

The rest follows exactly the same as before. Hence the proof is complete. □

B.1.4 Proof of Corollary 3.3.1

Proof. We prove the corollary by showing that the conditions in Theorem 3.3 are satisfied. In particular, it suffices to verify that for fixed $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, there exists a positive constant $C'_{0,\alpha}$ such that $\sup_{f_0, f_1 \in \mathcal{M}} \mathbb{P}_{f_0,f_1}(t_\alpha < C'_{0,\alpha} \delta_n) \geq 1 - \beta/2$. Then the rest of the proof proceeds the same as before.

• **i.i.d. sampling**

To start with the case of i.i.d. sampling, let $\eta = (\eta_1, \ldots, \eta_n)^\top$ be a permutation of $\{1, \ldots, n\}$. Now conditioned on the data $\mathcal{X}_{2n} = \{(X_1, Y_1), \ldots, (X_{2n}, Y_{2n})\}$, we denote the probability and expectation under permutations by $\mathbb{P}_\eta[\cdot] = \mathbb{P}_\eta[\cdot|\mathcal{X}_{2n}]$ and $\mathbb{E}_\eta[\cdot] = \mathbb{E}_\eta[\cdot|\mathcal{X}_{2n}]$ respectively. Then by Markov's inequality

$$\begin{aligned}\mathbb{P}_\eta\left(\widehat{\mathcal{T}}'_{global} \geq t\right) &= \mathbb{P}_\eta\left(\frac{1}{n}\sum_{i=n+1}^{2n}(\widehat{m}_\eta(X_i) - \widehat{\pi}_1)^2 \geq t\right) \\ &\leq \frac{1}{tn}\sum_{i=n+1}^{2n}\mathbb{E}_\eta\left[(\widehat{m}_\eta(X_i) - \widehat{\pi}_1)^2\right],\end{aligned}$$

where $\widehat{m}_\eta(x) = \sum_{i=1}^n w_i(x) Y_{\eta_i}$. Since $\sum_{i=1}^n w_i(x) = 1$ for any $x \in S$,

$$\mathbb{E}_\eta\left[\widehat{m}_\eta(x)\right] = \sum_{i=1}^n w_i(x)\mathbb{E}_\eta[Y_{\eta_i}] = \sum_{i=1}^n w_i(x)\widehat{\pi}_1 = \widehat{\pi}_1.$$

Further note that

$$\begin{aligned}\mathbb{E}_\eta\left[(\widehat{m}_\eta(x) - \widehat{\pi}_1)^2\right] &= \sum_{i_1=1}^n\sum_{i_2=1}^n w_{i_1}(x)w_{i_2}(x)\mathbb{E}_\eta\left[(Y_{\eta_{i_1}} - \widehat{\pi}_1)(Y_{\eta_{i_2}} - \widehat{\pi}_1)\right] \qquad \text{(B.3)} \\ &\leq \sum_{i=1}^n w_i^2(x)\mathbb{E}_\eta\left[(Y_{\eta_i} - \widehat{\pi}_1)^2\right] \\ &= \widehat{\pi}_1(1 - \widehat{\pi}_1)\sum_{i=1}^n w_i^2(x) \\ &\leq \frac{1}{4}\sum_{i=1}^n w_i^2(x), \qquad \text{(B.4)}\end{aligned}$$

where the first inequality uses $\mathbb{E}_\eta\left[(Y_{\eta_{i_1}} - \widehat{\pi}_1)(Y_{\eta_{i_2}} - \widehat{\pi}_1)\right] \leq 0$ when $i_1 \neq i_2$.

Note that the permutation samples are not *i.i.d.* and thus in order to use the condition in (3.9) which holds for *i.i.d.* samples, we will associate the upper bound in (B.4) with *i.i.d.* samples. To do so, let $(Y_1^*, \ldots, Y_n^*)$ be *i.i.d.* Bernoulli random variables with parameter $p = 1/2$ independent of $\{X_1, \ldots, X_{2n}\}$. Then

$$\begin{aligned}&\mathbb{E}_{Y^*}\left[(\widehat{m}(x) - 1/2)^2 | X_1, \ldots, X_{2n}\right] \\ =\ &\mathbb{E}_{Y^*}\Big[\big(\sum_{i=1}^n w_i(x)Y_i^* - 1/2\big)^2 \big| X_1, \ldots, X_{2n}\Big] \\ =\ &\mathbb{E}_{Y^*}\Big[\big(\sum_{i=1}^n w_i(x)(Y_i^* - 1/2)\big)^2 \big| X_1, \ldots, X_{2n}\Big]\end{aligned}$$

$$= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1}(x)w_{i_2}(x)\mathbb{E}_{Y^*}[(Y^*_{i_1}-1/2)(Y^*_{i_2}-1/2)]$$
$$= \frac{1}{4}\sum_{i=1}^{n} w_i^2(x).$$

Therefore, we obtain

$$\mathbb{E}_\eta\left[(\widehat{m}_\eta(x)-\widehat{\pi}_1)^2\right] \le \mathbb{E}_{Y^*}\left[(\widehat{m}(x)-1/2)^2|X_1,\ldots,X_{2n}\right]$$

which in turn implies that

$$\mathbb{P}_\eta\left(\widehat{\mathcal{T}}'_{global}\ge t\right) \le \frac{1}{tn}\sum_{i=n+1}^{2n}\mathbb{E}_{Y^*}\left[(\widehat{m}(X_i)-1/2)^2|X_1,\ldots,X_{2n}\right].$$

So the critical value of the permutation distribution is bounded by

$$t^*_\alpha \le \frac{1}{\alpha n}\sum_{i=n+1}^{2n}\mathbb{E}_{Y^*}\left[(\widehat{m}(X_i)-1/2)^2|X_1,\ldots,X_{2n}\right]. \tag{B.5}$$

Now choose $C'_{0,\alpha}$ such that $2C_0/(\alpha\beta)\le C'_{0,\alpha}$. Then based on the assumption in (3.9) and Markov's inequality

$$\sup_{f_0,f_1\in\mathcal{M}}\mathbb{P}_{f_0,f_1}\left(t^*_\alpha\ge C'_{0,\alpha}\delta_n\right)$$
$$\le \sup_{f_0,f_1\in\mathcal{M}}\mathbb{P}_{f_0,f_1}\left(\frac{1}{\alpha n}\sum_{i=n+1}^{2n}\mathbb{E}_{Y^*}\left[(\widehat{m}(X_i)-1/2)^2|X_1,\ldots,X_{2n}\right]\ge C'_{0,\alpha}\delta_n\right)$$
$$\le \frac{C_0}{C'_{0,\alpha}\alpha}\le\beta/2.$$

Hence the proof completes.

• **Separate Sampling**

Let $Y^{**}_1,\ldots,Y^{**}_n$ be Bernoulli random variables with parameter $\widehat{\pi}_1$ such that $\sum_{i=1}^n Y^{**}_i = n\widehat{\pi}_1$ and they are independent of $X_1,\ldots,X_{2n}$. In the case of separate sampling, the proof follows similarly by noting that the right-hand side of (B.3) is the same as

$$\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1}(x)w_{i_2}(x)\mathbb{E}_\eta\left[(Y_{\eta_{i_1}}-\widehat{\pi}_1)(Y_{\eta_{i_2}}-\widehat{\pi}_1)\right]$$
$$= \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} w_{i_1}(x)w_{i_2}(x)\mathbb{E}_{Y^{**}}\left[(Y^{**}_{i_1}-\widehat{\pi}_1)(Y^{**}_{i_2}-\widehat{\pi}_1)\right]$$

$$= \mathbb{E}_{Y^{**}}[(\widehat{m}(x) - \widehat{\pi}_1)^2 | X_1, \ldots, X_n].$$

Now by putting the above quantity into the right-hand side of (B.5) and following the same lines afterwards, we complete the proof. □

B.1.5 Proof of Theorem 3.4

This result can be proved by following the same steps in the proof of Theorem 3.3. In fact, it is simpler than the previous proof since it does not involve sample splitting to estimate the integration error; hence we omit the proof.

B.1.6 Proof of Example 3.1

Proof. Let $\overline{m}_{kNN}(x) = \mathbb{E}[\widehat{m}_{kNN}(x)|X_1, \ldots, X_n]$. Then we have the following decomposition.

$$\mathbb{E}\left[(\widehat{m}_{kNN}(x) - m(x))^2\right] = \underbrace{\mathbb{E}\left[(\widehat{m}_{kNN}(x) - \overline{m}_{kNN}(x))^2\right]}_{(I)} + \underbrace{\mathbb{E}\left[(\overline{m}_{kNN}(x) - m(x))^2\right]}_{(II)}.$$

For a fixed x, Proposition 8.1 of Biau and Devroye (2015) shows that conditioned on $\{X_1, \ldots, X_n\}$,

$$(X_{1,n}(x), Y_{1,n}(x)), \ldots, ((X_{n,n}(x), Y_{n,n}(x)))$$

are independent. Using this independence property,

$$(I) = \mathbb{E}\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n}(Y_{i,n}(x) - m(X_{i,n}(x)))\right)^2\right] \leq \frac{1}{4k_n}.$$

Next for (II),

$$\begin{aligned}
(II) &= \mathbb{E}\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n}(m(X_{i,n}(x)) - m(x))\right)^2\right] \\
&\leq \mathbb{E}\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\left|m(X_{i,n}(x)) - m(x)\right|\right)^2\right] \\
&\leq \mathbb{E}\left[\left(\frac{L}{k_n}\sum_{i=1}^{k_n}||X_{i,n}(x) - x||_2\right)^2\right]
\end{aligned}$$

where the last inequality uses the Lipschitz condition. Note that for fixed $\epsilon > 0$

$$\begin{aligned}\mathbb{P}\left(||X_{1,n}(x) - x||_2 > \epsilon\right) &= \left(1 - \mathbb{P}(X \in B_{x,\epsilon})\right)^n \\ &\leq (1 - \tau_x \epsilon^D)^n \leq \ e^{-\tau_x n \epsilon^D}\end{aligned} \tag{B.6}$$

by the assumption that $\mathbb{P}(X \in B_{x,\epsilon}) > \tau_x \epsilon^D$. Hence,

$$\begin{aligned}\mathbb{E}\left[||X_{1,n}(x) - x||^2\right] &= \int_0^\infty \mathbb{P}\left(||X_{1,n}(x) - x||_2 > \sqrt{\epsilon}\right) d\epsilon \\ &\leq \int_0^\infty e^{-\tau_x n \epsilon^{D/2}} d\epsilon \\ &= \frac{2\Gamma(2/D)}{D\tau_x^{2/D}} n^{-2/D}.\end{aligned} \tag{B.7}$$

Similarly to the proof of Theorem 6.2 of Györfi et al. (2002), divide the data into $k_n + 1$ parts where the first k_n parts have size $\lfloor n/k_n \rfloor$ and denote the first nearest neighbor of x from the jth partition by $\widetilde{X}_j^x$. This implies that

$$\sum_{i=1}^{k_n} ||X_{i,n}(x) - x||_2 \leq \sum_{i=1}^{k_n} ||\widetilde{X}_i^x - x||_2$$

and by Jensen's inequality,

$$\begin{aligned}(II) \leq \ \mathbb{E}\left[\left(\frac{L}{k_n}\sum_{i=1}^{k_n} ||\widetilde{X}_i^x - x||_2\right)^2\right] &\leq \ \frac{L^2}{k_n}\sum_{i=1}^{k_n}\mathbb{E}\left[||\widetilde{X}_i^x - x||_2^2\right] \\ &\leq \ L^2 \frac{2\Gamma(2/D)}{D\tau_x^{2/D}}\left(\frac{k_n}{n}\right)^{2/D}\end{aligned}$$

by the inequality (B.7). Combining the results, we have

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{m}_{kNN}(x) - m(x)\right)^2\right] &= \ (I) + (II) \\ &\leq \ \frac{1}{4k_n} + L^2 \frac{2\Gamma(2/D)}{D\tau_x^{2/D}}\left(\frac{k_n}{n}\right)^{2/D}.\end{aligned}$$

This completes the proof. □

B.1.7 Proof of Example 3.2

Proof. Following the proof of Example 3.1, let

$$\overline{m}_{ker}(x) = \mathbb{E}\left[\widehat{m}_{ker}(x)|X_1, \ldots, X_n\right]$$

and thus

$$\mathbb{E}\left[\left(\widehat{m}_{ker}(x) - m(x)\right)^2\right] = \underbrace{\mathbb{E}\left[\left(\widehat{m}_{ker}(x) - \overline{m}_{ker}(x)\right)^2\right]}_{(I)} + \underbrace{\mathbb{E}\left[\left(\overline{m}_{ker}(x) - m(x)\right)^2\right]}_{(II)}.$$

Define an event

$$\mathcal{A}_n = \left\{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \geq \lambda\right\}.$$

Then

$$(I) = \underbrace{\mathbb{E}\left[\left(\widehat{m}_{ker}(x) - \overline{m}_{ker}(x)\right)^2 I(\mathcal{A}_n)\right]}_{(I_1)} + \underbrace{\mathbb{E}\left[\left(\widehat{m}_{ker}(x) - \overline{m}_{ker}(x)\right)^2 I(\mathcal{A}_n^c)\right]}_{(I_2)}.$$

For (I_1), we have

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{m}_{ker}(x) - \overline{m}_{ker}(x)\right)^2 I(\mathcal{A}_n)|X_1, \ldots, X_n\right] &= \frac{\sum_{i=1}^{n} \mathrm{Var}(Y_i|X_i) K\left(\frac{x-X_i}{h_n}\right)}{\left(\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\right)^2} I(\mathcal{A}_n) \\
&\leq \frac{1}{4\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)} I\left(\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \geq \lambda\right) \\
&\leq \frac{1 + \lambda^{-1}}{4 + 4\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)} \\
&\leq \frac{1 + \lambda^{-1}}{4 + 4\lambda\sum_{i=1}^{n} I\left(||x - X_i||_2 \leq r h_n\right)} \\
&\leq \frac{1 + \lambda}{4\lambda^2} \frac{1}{1 + \sum_{i=1}^{n} I\left(||x - X_i||_2 \leq r h_n\right)}.
\end{aligned}$$

By Lemma 4.1 of Györfi et al. (2002),

$$\mathbb{E}\left[\frac{1}{1+B}\right] \leq \frac{1}{(n+1)p} \leq \frac{1}{np},$$

where $B \sim \text{Binominal}(n, p)$. Using this result,

$$(I_1) \leq \frac{1+\lambda}{4\lambda^2} \frac{1}{n\mathbb{P}\left(X \in B_{x,rh_n}\right)} \leq \left(\frac{1+\lambda}{4\lambda^2 \tau_x r^d}\right) \frac{1}{nh_n^d}.$$

For (I_2), note that $\left(\widehat{m}_{ker}(x) - \overline{m}_{ker}(x)\right)^2 \leq 1$ and thus

$$\begin{aligned}(I_2) \leq & \ \mathbb{P}\left(\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) < \lambda\right) \\ \leq & \ \mathbb{P}\left(\sum_{i=1}^n I\left(||x-X_i||_2 \leq rh_n\right) = 0\right)\end{aligned}$$

where the second inequality is because if there exists X_i such that $||x-X_i||_2 \leq rh_n$, then $\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) \geq \lambda$ by the assumption on the kernel. In addition,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n I\left(||x-X_i||_2 \leq rh_n\right) = 0\right) = & \ \left(1 - \mathbb{P}\left(X \in B_{x,rh_n}\right)\right)^n \\ \overset{(i)}{\leq} & \ e^{-n\tau_x r^D h_n^D} \overset{(ii)}{\leq} \left(\frac{e^{-1}}{\tau_x r^D}\right) \frac{1}{nh_n^D},\end{aligned} \tag{B.8}$$

where (i) uses $1 + x \leq e^x$ with the assumption $\mathbb{P}\left(X \in B_{x,\epsilon}\right) \geq \tau_x \epsilon^D$ and (ii) uses $\sup_z ze^{-z} \leq e^{-1}$. As a result,

$$(I) = (I_1) + (I_2) \leq \left(\frac{1+\lambda}{4\lambda^2 \tau_x r^D} + \frac{e^{-1}}{\tau_x r^D}\right) \frac{1}{nh_n^D}.$$

For (II), we use Jensen's inequality and the Lipschitz condition to have

$$\begin{aligned}& \left(\overline{m}_{ker}(x) - m_{ker}(x)\right)^2 \\ = & \left(\frac{\sum_{i=1}^n \left(m(X_i) - m(x)\right) K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)}\right)^2 I\left(\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) > 0\right) + m_{ker}(x)^2 I\left(\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) = 0\right) \\ \leq & \frac{\sum_{i=1}^n L^2 ||X_i - x||_2^2 K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} I\left(\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) > 0\right) + I\left(\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right) = 0\right).\end{aligned}$$

Since $K(x) \leq I(x \in B_{0,R})$, we observe that

$$||X_i - x||_2^2 K\left(\frac{x-X_i}{h_n}\right) \leq R^2 h_n^2 K\left(\frac{x-X_i}{h_n}\right).$$

Consequently,

$$\begin{aligned}\left(\overline{m}_{ker}(x) - m_{ker}(x)\right)^2 \leq\ & L^2R^2h_n^2 + I\left(\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) = 0\right) \\ \leq\ & L^2R^2h_n^2 + I\left(\sum_{i=1}^{n} I\left(||x - X_i||_2 \leq rh_n\right) = 0\right),\end{aligned}$$

where the second inequality is by the assumption $\lambda I(x \in B_{0,r}) \leq K(x)$. By taking the expectation,

$$\begin{aligned}(II) \leq\ & L^2R^2h_n^2 + \left(1 - \mathbb{P}\left(X \in B_{x,rh_n}\right)\right)^n \\ \leq\ & L^2R^2h_n^2 + \left(1 - \tau_x r^D h_n^D\right)^n \\ \leq\ & L^2R^2h_n^2 + \left(\frac{e^{-1}}{\tau_x r^D}\right)\frac{1}{nh_n^D}.\end{aligned} \tag{B.9}$$

Therefore, we conclude that

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{m}_{ker}(x) - m(x)\right)^2\right] =\ & (I) + (II) \\ \leq\ & \left(\frac{1+\lambda}{4\lambda^2\tau_x r^D} + \frac{2e^{-1}}{\tau_x r^D}\right)\frac{1}{nh_n^D} + L^2R^2h_n^2,\end{aligned}$$

which completes the proof. □

B.1.8 Proof of Theorem 4.8

Proof. Suppose X has the uniform distribution over $[0, B]^D$ and $B > 0$. In addition, assume that for $0 < \epsilon < 1/2$, the regression function is given by

$$\begin{aligned}m(x) =\ & \epsilon\prod_{i=1}^{D}\left(1 - \frac{x_i}{B\epsilon}\right) I\left(0 \leq x_i \leq B\epsilon\right) \\ & + \epsilon\prod_{i=1}^{D}\left(\frac{B(1-\epsilon) - x_i}{B\epsilon}\right) I\{B(1-\epsilon) \leq x_i \leq B\} + \frac{1}{2}\end{aligned} \tag{B.10}$$

for $x = (x_1, \ldots, x_D) \in [0, B]^D$ and $m(x) = 0$ otherwise. Therefore, we have $\pi_1 = \pi_0 = 1/2$. Now for any $x, z \in [0, B]^D$, the telescoping argument gives

$$\begin{aligned}& |m(x_1, \ldots, x_D) - m(z_1, \ldots, z_D)| \\ \leq\ & |m(x_1, x_2, \ldots, x_D) - m(z_1, x_2, \ldots, x_D)|\end{aligned}$$

$$+ \sum_{i=1}^{D-2} |m(z_1, \ldots, z_i, x_{i+1}, \ldots, x_D) - m(z_1, \ldots, z_i, z_{i+1}, x_{i+2}, \ldots, x_D)|$$

$$+ |m(z_1, z_2, \ldots, z_{D-1}, x_D) - m(z_1, z_2, \ldots, z_D)|.$$

For the first term,

$$\begin{aligned}
&|m(x_1, x_2, \ldots, x_D) - m(z_1, x_2, \ldots, x_D)| \\
\leq\ & \epsilon \left| \left(1 - \frac{x_1}{B\epsilon}\right) I\left(0 \leq x_1 \leq B\epsilon\right) - \left(1 - \frac{z_1}{B\epsilon}\right) I\left(0 \leq z_1 \leq B\epsilon\right) \right| \times \prod_{i=2}^{D} \left| \left(1 - \frac{x_i}{B\epsilon}\right) I\left(0 \leq x_i \leq B\epsilon\right) \right| \\
&+\epsilon \left| \left(\frac{B(1-\epsilon) - x_1}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq x_1 \leq B\Big\} - \left(\frac{B(1-\epsilon) - z_1}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq z_1 \leq B\Big\} \right| \\
&\quad \times \prod_{i=2}^{D} \left| \left(\frac{B(1-\epsilon) - x_i}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq x_i \leq B\Big\} \right| \\
\leq\ & \epsilon \left| \left(1 - \frac{x_1}{B\epsilon}\right) I\left(0 \leq x_1 \leq B\epsilon\right) - \left(1 - \frac{z_1}{B\epsilon}\right) I\left(0 \leq z_1 \leq B\epsilon\right) \right| \\
&+\epsilon \left| \left(\frac{B(1-\epsilon) - x_1}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq x_1 \leq B\Big\} - \left(\frac{B(1-\epsilon) - z_1}{B\epsilon}\right) I\Big\{B(1-\epsilon) \leq z_1 \leq B\Big\} \right| \\
\leq\ & \frac{2}{B}|x_1 - z_1| \leq \frac{2}{B}||x - z||_2.
\end{aligned}$$

Applying the same logic to the other terms, we see that

$$|m(x) - m(z)| \leq \frac{2D}{B}||x - z||_2.$$

By choosing $B = 2D/L$, the regression function $m(x)$ becomes L-Lipschitz with

$$\delta_{n,x} = |m(x) - \pi_1|^2 = \epsilon^2 \quad \text{at } x = (0, \ldots, 0).$$

Next, we lower bound the testing error. Denote the product and joint measure of (X, Y) described above by P_0 and P_1 respectively. Using the standard approach to lower bound the testing error (e.g. Baraud, 2002), we obtain that for any α level test functions $\phi : \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \mapsto \{0, 1\}$,

$$\inf_{\phi \in \Phi_{n,\alpha}} \sup_{f_0, f_1 \in \mathcal{M}_{Lip}(\delta_{n,x})} \mathbb{P}_{f_0, f_1}(\phi = 0) \geq\ 1 - \alpha - \mathrm{TV}(P_0^n, P_1^n)$$

where TV denotes total variation distance. Based on Pinsker's inequality, we get

$$\mathrm{TV}(P_0^n, P_1^n) \leq \sqrt{\frac{n}{2} D_{KL}(P_1||P_0)}$$

where D_{KL} is the Kullback-Leibler divergence and by the Jensen's inequality

$$\begin{aligned}
& D_{KL}(P_1||P_0) \\
= & \int \pi_1 f(x) \log \frac{f(x, Y=1)}{\pi_1 f(x)} dx + \int (1-\pi_1) f(x) \log \frac{f(x, Y=0)}{(1-\pi_1) f(x)} dx \\
= & \frac{1}{2} \int f(x) \log \frac{f(x|Y=1)}{f(x)} dx + \frac{1}{2} \int f(x) \log \frac{f(x|Y=0)}{f(x)} dx \\
\leq & \frac{1}{2} \int \frac{(f(x|Y=1) - f(x))^2}{f(x)} dx + \frac{1}{2} \int \frac{(f(x|Y=0) - f(x))^2}{f(x)} dx.
\end{aligned}$$

By the assumption on (X, Y), X has the marginal density $f(x) = B^{-D}$ and the conditional densities $f(x|Y=1) = 2B^{-D} m(x)$ and $f(x|Y=0) = 2B^{-D} - f(x|Y=1)$ for $x \in [0, B]^D$. Therefore,

$$\begin{aligned}
& \frac{1}{2} \int \frac{(f(x|Y=1) - f(x))^2}{f(x)} dx + \frac{1}{2} \int \frac{(f(x|Y=0) - f(x))^2}{f(x)} dx \\
= & \int \frac{(f(x|Y=1) - f(x))^2}{f(x)} dx \\
= & 4B^{-D} \int (m(x) - 1/2)^2 \, dx.
\end{aligned}$$

Using the definition of $m(x)$ in (B.10), the above integration is calculated by

$$4B^{-D} \int (m(x) - 1/2)^2 \, dx = \frac{8}{3^D} \epsilon^{2+D}.$$

Now by choosing $\epsilon = \beta^{2/(2+D)} 3^{D/(2+D)} 2^{-2/(2+D)} n^{-1/(2+D)}$, we have

$$\inf_{\phi \in \Phi_{n,\alpha}} \sup_{f_0, f_1 \in \mathcal{M}_{Lip}(C_{1,x} n^{-2/(2+D)})} \mathbb{P}_{f_0, f_1}(\phi = 0) \geq 1 - \alpha - \beta.$$

This completes the proof. □

B.1.9 Proof of Proposition 3.1

It is enough to show that there exist universal constants $C_0, C'_{0,\alpha}$ such that

$$\sup_{f_0, f_1 \in \mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{kNN}(x) - m(x))^2 \right] \leq C_0 n^{-\frac{2}{2+d}},$$

$$\sup_{f_0,f_1\in\mathcal{M}_{Lip}} \mathbb{E}\left[(\widehat{m}_{ker}(x)-m(x))^2\right] \leq C'_{0,\alpha} n^{-\frac{2}{2+d}}.$$

Then we can apply Theorem 3.4 to complete the proof. To start with kNN regression, we only need to modify (B.6) and follow the same steps in the proof of Example 3.1. From the definition of the (C,d)-homogeneous measure, we see that

$$\mathbb{P}\left(X\in B_{x,\epsilon}\right) \geq \frac{\epsilon^d}{C}\mathbb{P}\left(X\in B_{x,1}\right) = C'\epsilon^d.$$

As a result, (B.6) becomes

$$\begin{aligned}\mathbb{P}\left(||X_{1,n}(x)-x||_2>\epsilon\right) = & \left(1-\mathbb{P}(X\in B_{x,\epsilon})\right)^n \\ \leq & \left(1-C'\epsilon^d\right)^n \leq e^{-C'n\epsilon^d}.\end{aligned}$$

Then we end up having

$$\mathbb{E}\left[(\widehat{m}_{kNN}(x)-m(x))^2\right] \leq \frac{1}{4k_n} + L^2\frac{2\Gamma(2/d)}{dC'^{2/d}}\left(\frac{k_n}{n}\right)^{2/d}$$

and the result follows by setting $k_n = n^{\frac{2}{2+d}}$. Similarly, we only need to modify (B.8) and (B.9) in the proof of Example 3.2. By using the (C,d)-homogeneous measure,

$$\begin{aligned}\left(1-\mathbb{P}\left(X\in B_{x,rh_n}\right)\right)^n \leq & \left(1-\frac{h_n^d}{C}\mathbb{P}\left(X\in B_{x,r}\right)\right)^n \\ = & \left(1-C'h_n^d\right)^n \\ \leq & e^{-C'nh_n^d}\end{aligned}$$

and apply this result to (B.8) and (B.9). We complete the proof by following the same steps in the proof of Example 3.2.

B.1.10 Proof of Theorem 3.6

Proof. We use a combinatorial central limit theorem in Bolthausen (1984) to prove the result. First denote $a_{ij} = w_i(x)Y_j$ for $1\leq i,j\leq n$ and

$$\mu = na_{\cdot\cdot}, \quad \sigma_n^2 = \sum_{1\leq i,j\leq n}(a_{ij}-a_{i\cdot}-a_{\cdot j}+a_{\cdot\cdot})^2/(n-1),$$

where

$$a_{i\cdot} = \sum_{j=1}^{n} a_{ij}/n, \quad a_{\cdot j} = \sum_{i=1}^{n} a_{ij}/n, \quad a_{\cdot\cdot} = \sum_{1\le i,j\le n}^{n} a_{ij}/n^2.$$

In our case, $\mu = \widehat{\pi}_1$ and σ_n^2 is given in (3.18). Let $d_{ij} = a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot} = (w_i(x) - 1/n)(Y_j - \widehat{\pi}_1)$. Then using the theorem in Bolthausen (1984), we obtain

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\widehat{m}(x) - \widehat{\pi}_1}{\sigma_n} \le t \middle| X_1, \ldots, X_n \right) - \Phi(t) \right| \le K \frac{1}{\sqrt{n}} \frac{\frac{1}{n^2}\sum_{i,j} |d_{i,j}|^3}{\left(\frac{1}{n^2}\sum_{i,j} d_{i,j}^2\right)^{3/2}},$$

where K is a universal constant. Note that

$$\frac{1}{n^2}\sum_{i,j} |d_{i,j}|^3 = \frac{1}{n}\sum_{i=1}^{n} \left| w_i(x) - \frac{1}{n} \right|^3 \cdot \frac{1}{n}\sum_{j=1}^{n} \left| Y_j - \widehat{\pi}_1 \right|^3$$

and

$$\frac{1}{n^2}\sum_{i,j} d_{i,j}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(w_i(x) - \frac{1}{n} \right)^2 \cdot \frac{1}{n}\sum_{j=1}^{n} \left(Y_j - \widehat{\pi}_1 \right)^2.$$

As a result,

$$\begin{aligned} \frac{\frac{1}{n^2}\sum_{i,j} |d_{i,j}|^3}{\left(\frac{1}{n^2}\sum_{i,j} d_{i,j}^2\right)^{3/2}} &= \frac{1}{\sqrt{n}} \frac{\frac{1}{n}\sum_{i=1}^{n} \left| w_i(x) - \frac{1}{n} \right|^3}{\left\{ \frac{1}{n}\sum_{i=1}^{n} \left(w_i(x) - \frac{1}{n} \right)^2 \right\}^{3/2}} \cdot \underbrace{\frac{\frac{1}{n}\sum_{j=1}^{n} \left| Y_j - \widehat{\pi}_1 \right|^3}{\left(\frac{1}{n}\sum_{j=1}^{n} \left(Y_j - \widehat{\pi}_1 \right)^2 \right)^{3/2}}}_{(II)} \\ &\le \underbrace{\frac{\max_{1\le i\le n} (w_i(x) - 1/n)^2}{\sum_{i=1}^{n} (w_i(x) - 1/n)^2}}_{(I)} \cdot (II) \end{aligned}$$

Note that $(I) = o_P(1)$ under the given assumption and (II) is stochastically bounded by the law of large number. Thus we conclude that

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\widehat{m}(x) - \widehat{\pi}_1}{\sigma_n} \le t \middle| X_1, \ldots, X_n \right) - \Phi(t) \right| = o_P(1),$$

which implies the desired result. □

B.1.11 Proof of Corollary 3.6.1

Proof. For kNN regression, there are k and $(n-k)$ number of k^{-1} and zero in $\{w_1(x), \ldots, w_n(x)\}$ respectively. Hence,

$$\sum_{i=1}^{n} \left(w_i(x) - \frac{1}{n} \right)^2 = k \left(\frac{1}{k} - \frac{1}{n} \right)^2 + \frac{n-k}{n^2}.$$

Furthermore, under the assumption that $2k < n$, we have

$$\max_{1 \le i \le n} \left| w_i(x) - \frac{1}{n} \right| = \frac{1}{k} - \frac{1}{n}.$$

After direct calculations, one can show that

$$\frac{\max_{1 \le i \le n} |w_i(x) - 1/n|}{\{\sum_{i=1}^{n} (w_i(x) - 1/n)^2\}^{1/2}} \to 0,$$

and thus the result follows. □

B.1.12 Proof of Corollary 3.6.2

Proof. Note that

$$\widehat{m}_{ker}(x) = \sum_{i=1}^{n} w_i(x) Y_i = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)} = \frac{\sum_{i=1}^{n} Y_i K_{h_n}(x - X_i)}{\sum_{i=1}^{n} K_{h_n}(x - X_i)}.$$

Hence it suffices to show that

$$\begin{aligned} & \frac{\max_{1 \le i \le n} (w_i(x) - 1/n)^2}{\sum_{i=1}^{n} (w_i(x) - 1/n)^2} \\ = \; & \frac{\max_{1 \le i \le n} \left(K_h(x - X_i) - \frac{1}{n} \sum_{j=1}^{n} K_h(x - X_j) \right)^2}{\sum_{i=1}^{n} \left(K_h(x - X_i) - \frac{1}{n} \sum_{j=1}^{n} K_h(x - X_j) \right)^2} \xrightarrow{p} 0. \end{aligned}$$

Using the given condition, the numerator is bounded by

$$\max_{1 \le i \le n} \left(K_h(x - X_i) - \frac{1}{n} \sum_{j=1}^{n} K_h(x - X_j) \right)^2 \le 4 h^{-D} \mathcal{K}^2.$$

Whereas the denominator can be decomposed into two parts:

$$\sum_{i=1}^{n}\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^{n}K_h(x-X_j)\right)^2 = \sum_{i=1}^{n}K_h^2(x-X_i)-2n\left(\frac{1}{n}\sum_{j=1}^{n}K_h(x-X_j)\right)^2$$

Based on the usual bias-variance decomposition of the kernel density estimation (Wasserman, 2006), each part can be approximated as

$$\frac{1}{nh^D}\sum_{i=1}^{n}K^2\left(\frac{x-X_i}{h}\right) = f(x)\int K^2(u)du + O(h) + O_P\left(\frac{1}{\sqrt{nh^D}}\right)$$
$$\frac{1}{nh^D}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) = f(x) + O(h^2) + O_P\left(\frac{1}{\sqrt{nh^D}}\right).$$

Now, the sufficient condition can be further bounded by

$$\begin{aligned}
&\frac{\max_{1\le i\le n}\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^{n}K_h(x-X_j)\right)^2}{\sum_{i=1}^{n}\left(K_h(x-X_i)-\frac{1}{n}\sum_{j=1}^{n}K_h(x-X_j)\right)^2}\\
\le\ &\frac{4h^{-D}\mathcal{K}^2}{\frac{1}{h^{2D}}\sum_{i=1}^{n}K^2\left(\frac{x-X_i}{h}\right)-2n\left(\frac{1}{nh^D}\sum_{j=1}^{n}K\left(\frac{x-X_j}{h}\right)\right)^2}\\
=\ &\frac{4n^{-1}\mathcal{K}^2}{\frac{1}{nh^D}\sum_{i=1}^{n}K^2\left(\frac{x-X_i}{h}\right)-2h^D\left(\frac{1}{nh^D}\sum_{j=1}^{n}K\left(\frac{x-X_j}{h}\right)\right)^2}. \qquad \text{(B.11)}
\end{aligned}$$

Then using the previous approximations, the denominator becomes

$$\begin{aligned}
&\frac{1}{nh^D}\sum_{i=1}^{n}K^2\left(\frac{x-X_i}{h}\right)-2h^D\left(\frac{1}{nh^D}\sum_{j=1}^{n}K\left(\frac{x-X_j}{h}\right)\right)^2\\
=\ &f(x)\int K^2(u)du + O(h) + O_P\left(\frac{1}{\sqrt{nh^D}}\right) - 2h^D\left(f(x)+O(h^2)+O_P\left(\frac{1}{\sqrt{nh^D}}\right)\right)^2\\
=\ &\underbrace{f(x)\int K^2(u)du}_{>0 \text{ by the assumption}} + o_P(1).
\end{aligned}$$

Hence (B.11) converges to zero in probability and the result follows. □

B.2 Diffusion Maps

Dimensionality reduction methods can be useful for visualizing and describing low-dimensional structures that are embedded in higher-dimensional spaces. In this section, we briefly describe diffusion maps (Coifman et al., 2005; Coifman and Lafon, 2006) and the particular version that we use to visualize the results of our local two-sample test.

As a starting point for constructing a diffusion map, one first defines a weight that reflects the local similarity of two points x_i and x_j in $\mathcal{X} = \{x_1, \dots, x_n\}$. A common choice is the Gaussian kernel

$$w(x_i, x_j) = \exp\left(-\frac{s(x_i, x_j)^2}{\epsilon}\right), \tag{B.12}$$

where $s(x_i, x_j)$ represents (for example, the Euclidean) distance between the points. These weights are used to build a Markov random walk on the data with the transition probability from x_i to x_j defined as

$$p(x_i, x_j) = \frac{w(x_i, x_j)}{\sum_{k \in \Omega} w(x_i, x_k)}.$$

The one-step transition probabilities are stored in an $n \times n$ matrix denoted by $\mathbf{P}$, and then usually propagated by a t-step Markov random walk with transition probabilities $\mathbf{P}^t$. Instead of choosing a fixed time parameter t, however, we here combine diffusions at all times (Coifman et al., 2005) and define an averaged diffusion map* according to

$$\Psi_{\text{av}} : x \mapsto \left[\left(\frac{\lambda_1}{1-\lambda_1}\right)\psi_1(x), \left(\frac{\lambda_2}{1-\lambda_2}\right)\psi_2(x), \dots, \left(\frac{\lambda_m}{1-\lambda_m}\right)\psi_m(x)\right],$$

where λ_i and ψ_i, respectively, represent the first mth eigenvalues and the corresponding right eigenvectors of $\mathbf{P}$.

In our application for galaxy morphologies, we also use a generalization of the weight in (B.12) proposed by Zelnik-Manor and Perona (2005) for spectral clustering. In their paper, the authors show that a data-driven varying bandwidth leads to more meaningful clustering results for data with multiple scales and propose the weight

$$\widehat{w}(x_i, x_j) = \exp\left(-\frac{s(x_i, x_j)^2}{\sigma_i \sigma_j}\right),$$

where $\sigma_{i(j)}$ is the distance between $x_{i(j)}$ and its kth neighbor. For our visualization purposes, we choose $m = 2$ and $k = 50$, but a range of other values give similar results.

*This is also the default option of the function `diffuse()` in the `R` package `diffusionMap`.

Appendix C

Appendix for Chapter 4

C.1 Outline

This chapter is organized as follows:

- Appendix C.2 presents the asymptotic behavior of the permutation distribution of a two-sample degenerate U-statistic.
- Appendix C.3 collects several auxiliary lemmas, based on which we prove the main results of this paper.
- Appendix C.4 contains all proofs for the results in the main text.
- Appendix C.5 collects some details omitted in the main text as well as further applications of projection-averaging.
- Appendix C.6 contains additional simulation results both under high-dimensional and low-dimensional settings.

C.2 Permutation Tests

In this section, we study the limiting behavior of the permutation distribution of a two-sample U-statistic under the conventional asymptotic framework (4.5). Specifically, we establish fairly general conditions under which the permutation distribution of a two-sample U-statistic is asymptotically equivalent to the corresponding unconditional null distribution. We first focus on the large sample behavior of the permutation distribution under the null hypothesis in Section C.2.1 and then discuss how to generalize this result to the alternative hypothesis via coupling argument in Section C.2.2.

C.2.1 Asymptotic null behavior of permutation U-statistics

Let us start with some notation. For $r \geq 2$, consider a kernel $g(x_1, \ldots, x_r; y_1, \ldots, y_r)$ of degree (r, r) such that

$$\begin{aligned} &\mathbb{E}\left[g(X_1, \ldots, X_r; Y_1, \ldots, Y_r)\right] = \theta, \\ &\mathbb{E}\left[\{g(X_1, \ldots, X_r; Y_1, \ldots, Y_r)\}^2\right] < \infty. \end{aligned} \tag{C.1}$$

Without loss of generality, we assume that $g(x_1, \ldots, x_r; y_1, \ldots, y_r)$ is symmetric in each set of arguments, which means that the value of the kernel is invariant to the order of the first r arguments as well as the last r arguments. The reason for this is that we can always redefine the kernel as

$$\begin{aligned} &\widetilde{g}(x_1, \ldots, x_r; y_1, \ldots, y_r) \\ &= \frac{1}{r!r!} \sum_{\varpi \in \mathcal{S}_r} \sum_{\varpi' \in \mathcal{S}_r} g(x_{\varpi(1)}, \ldots, x_{\varpi(r)}; y_{\varpi'(1)}, \ldots, y_{\varpi'(r)}), \end{aligned} \tag{C.2}$$

where $\mathcal{S}_r$ is the set of all permutations of $\{1, \ldots, r\}$.

Let us write the U-statistic based on the kernel g by

$$U_{m,n} = \binom{m}{r}^{-1} \binom{n}{r}^{-1} \sum_{\alpha_1, \ldots, \alpha_r} \sum_{\beta_1, \ldots, \beta_r} g(X_{\alpha_1}, \ldots, X_{\alpha_r}; Y_{\beta_1}, \ldots, Y_{\beta_r}), \tag{C.3}$$

where the sums are taken over all subsets $\{\alpha_1, \ldots, \alpha_r\}$ of $\{1, \ldots, m\}$ and $\{\beta_1, \ldots, \beta_r\}$ of $\{1, \ldots, n\}$ and $\binom{m}{r}$ and $\binom{n}{r}$ are the binomial coefficient defined by $m!/\{r!(m-r)!\}$ and $n!/\{r!(n-r)!\}$, respectively. For $0 \leq c, d \leq r$, let $g_{c,d}(x_1, \ldots, x_c; y_1, \ldots, y_d)$ be the conditional expectation given by

$$\begin{aligned} &g_{c,d}(x_1, \ldots, x_c; y_1, \ldots, y_d) \\ &:= \mathbb{E}\big[g(x_1, \ldots, x_c, X_{c+1}, \ldots, X_r; y_1, \ldots, y_d, Y_{d+1}, \ldots, Y_r)\big]. \end{aligned} \tag{C.4}$$

Further write the centered conditional expectation and its variance as

$$g^*_{c,d}(x_1, \ldots, x_c; y_1, \ldots, y_d) := g_{c,d}(x_1, \ldots, x_c; y_1, \ldots, y_d) - \theta, \tag{C.5}$$

$$\sigma^2_{c,d} := \mathrm{Var}\left[g_{c,d}(X_1, \ldots, X_c; Y_1, \ldots, Y_d)\right] = \mathbb{E}\big[\big\{g^*_{c,d}(X_1, \ldots, X_c; Y_1, \ldots, Y_d)\big\}^2\big]. \tag{C.6}$$

The kernel g is *non-degenerate* if both $\sigma_{0,1}$ and $\sigma_{1,0}$ are strictly positive, and *degenerate* if $\sigma_{0,1} = \sigma_{1,0} = 0$. For the case where the kernel is non-degenerate, Chung and Romano (2016a) provided a sufficient condition under which the permutation distribution approximates the unconditional distribution of $U_{m,n}$. Their result, however, does not cover some important degenerate U-statistics including U_{CvM}, U_{Energy} and U_{MMD} in the main text. To fill this gap, we develop a similar result for the degenerate cases.

Consider the centered U-statistic scaled by $N = m + n$:

$$U^*_{m,n}(X_1, \ldots, X_m, Y_1, \ldots, Y_n) := N(U_{m,n} - \theta),$$

and let $\{Z_1, \ldots, Z_{m+n}\} = \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ be the pooled samples. Then the permutation distribution function of $U^*_{m,n}$ can be written as

$$\widehat{R}_{m,n}(t) = \frac{1}{N!} \sum_{\varpi \in \mathcal{S}_N} I\big\{U^*_{m,n}(Z_{\varpi(1)}, \ldots, Z_{\varpi(N)}) \leq t\big\}.$$

Also, let $R(t)$ be the unconditional limiting null distribution of $U^*_{m,n}$. Then we present the following theorem.

Theorem C.1 (Limiting behavior of the permutation distribution)**.** *Suppose $g(x_1, \ldots, x_r; y_1, \ldots, y_r)$ is symmetric in each set of arguments and degenerate under H_0. Further assume that $\mathbb{E}[g^2] < \infty$ and it satisfies*

Condition 1. $g^*_{0,2}(z_1, z_2) = g^*_{2,0}(z_1, z_2)$ *and* $g^*_{1,1}(z_1, z_2) = \frac{1-r}{r} g^*_{0,2}(z_1, z_2)$,

Condition 2. $\sigma^2_{0,1} = \sigma^2_{1,0} = 0$ *and* $\sigma^2_{0,2},\ \sigma^2_{2,0},\ \sigma^2_{1,1} > 0$,

Then under the conventional limiting regime (4.5) and H_0,

$$\sup_{t \in \mathbb{R}} \left| \widehat{R}_{m,n}(t) - R(t) \right| \xrightarrow{p} 0. \tag{C.7}$$

Proof. The proof can be found in Section C.4.23. □

C.2.2 The coupling argument

The proof of Theorem C.1 relies on the fact that $Z_{\varpi(1)}, \ldots, Z_{\varpi(N)}$ are *i.i.d.* samples under the null hypothesis for any permutations. The main difficulty of generalizing this result to the alternative hypothesis is that the given samples are not identically distributed under H_1. We instead have m samples $\{X_1, \ldots, X_m\}$ from P_X and n samples $\{Y_1, \ldots, Y_n\}$ from P_Y. In order to overcome such difficulty, we employ the coupling argument considered in Chung and Romano (2013), which is summarized in Algorithm 4.

Note that the output of Algorithm 4 consists of *i.i.d.* samples from $\vartheta_X P_X + \vartheta_Y P_Y$. We also note that there are $D = |m - B|$ different observations between the original samples $\{Z_1, \ldots, Z_N\}$ and the coupled samples $\{\overline{Z}_{\varpi_0(1)}, \ldots,$
$\overline{Z}_{\varpi_0(N)}\}$. The main strategy of studying the permutation distribution under the alternative hypothesis is to

Algorithm 4: Coupling

Data: $\{Z_1, \ldots, Z_N\} := \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ where $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} P_X$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} P_Y$, a random permutation ϖ_0 of $\{1, \ldots, N\}$.

Result: $\{\overline{Z}_{\varpi_0(1)}, \ldots, \overline{Z}_{\varpi_0(N)}\}$.

begin
- $B \sim \text{Binomial}(N, \vartheta_X)$;
- **if** $B \geq m$ **then**
  - Generate $\{X_{m+1}, \ldots, X_B\}$ *i.i.d.* samples from P_X;
  - **return** $\{\overline{Z}_{\varpi_0(1)}, \ldots, \overline{Z}_{\varpi_0(N)}\} := \{X_1, \ldots, X_m, Y_1, \ldots, Y_{N-B}, X_{m+1}, \ldots, X_B\}$;
- **else**
  - Generate $\{Y_{n+1}, \ldots, Y_{N-B}\}$ *i.i.d.* samples from P_Y;
  - **return** $\{\overline{Z}_{\varpi_0(1)}, \ldots, \overline{Z}_{\varpi_0(N)}\} := \{X_1, \ldots, X_B, Y_{n+1}, \ldots, Y_{N-B}, Y_1, \ldots, Y_n\}$;

establish that

$$U^*_{m,n}(Z_{\varpi(1)}, \ldots, Z_{\varpi(N)}) - U^*_{m,n}(\overline{Z}_{\varpi(\varpi_0(1))}, \ldots, \overline{Z}_{\varpi(\varpi_0(N))}) \xrightarrow{p} 0. \tag{C.8}$$

If this is the case, then both statistics have the same limiting behavior, which means that we can still apply Theorem C.1. We demonstrate this procedure by using the proposed CvM-statistic and prove Theorem 4.5 in the main text. The details can be found in the proof of Theorem 4.5.

C.3 Auxiliary Lemmas

In this section, we collect some auxiliary lemmas used in our main proofs. We start with another expression for the CvM-distance in Lemma C.1.1 and for the CvM-statistic in Lemma C.1.2. The variance of a two-sample U-statistic is given in Lemma C.1.3, which is useful to study the robustness of the CvM test in Theorem 4.7 and the minimax separation in Theorem 5.5. We recall Hoeffding's condition in Lemma C.1.4. Lemma C.1.5 extends the result of Chikkagoudar and Bhat (2014) to a multivariate case and studies the limiting behavior of a degenerate U-statistic under the contiguous alternative. In Lemma C.1.6 and Lemma C.1.7, we provide lower bounds for the CvM-distance that are used to prove Theorem 4.8. Lastly, Lemma C.1.8 generalizes Lemma 4.0.2 with three indicator functions.

Lemma C.1.1 (Another expression for the CvM-distance)**.** *Let* $X_1, X_2, X_3 \overset{i.i.d.}{\sim} P_X$ *and, independently,* $Y_1, Y_2, Y_3 \overset{i.i.d.}{\sim} P_Y$*. Furthermore, assume that* $\beta^\top X_1$ *and* $\beta^\top Y_1$ *have continuous distribution functions for* λ*-almost all* $\beta \in \mathbb{S}^{d-1}$*. Then the squared multivariate CvM-distance can be written as*

$$W_d^2(P_X, P_Y) \;=\; \frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}\left(X_1 - X_2, Y_1 - X_2\right)\right] + \frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}\left(X_1 - Y_2, Y_1 - Y_2\right)\right]$$

$$- \frac{1}{4\pi}\mathbb{E}\left[\mathsf{Ang}\left(X_1 - X_3, X_2 - X_3\right)\right] - \frac{1}{4\pi}\mathbb{E}\left[\mathsf{Ang}\left(X_1 - Y_1, X_2 - Y_1\right)\right]$$
$$- \frac{1}{4\pi}\mathbb{E}\left[\mathsf{Ang}\left(Y_1 - Y_3, Y_2 - Y_3\right)\right] - \frac{1}{4\pi}\mathbb{E}\left[\mathsf{Ang}\left(Y_1 - X_1, Y_2 - X_1\right)\right].$$

Proof. Since the CvM-distance is invariant to the choice of ϑ_X and ϑ_Y (Theorem 4.1), we may assume that $\vartheta_X = \vartheta_Y = 1/2$ for simplicity. Then

$$\begin{aligned} W_d^2 &= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(F_{\beta^\top X}(t) - F_{\beta^\top Y}(t)\right)^2 d\{F_{\beta^\top X}(t)/2 + F_{\beta^\top Y}(t)/2\} d\lambda(\beta) \\ &= \mathbb{E}\left[\left(F_{\beta^\top X}(\beta^\top Z^*)\right)^2\right] + \mathbb{E}_{\beta, Z^*}\left[\left(F_{\beta^\top Y}(\beta^\top Z^*)\right)^2\right] - 2\mathbb{E}\left[F_{\beta^\top X}(\beta^\top Z^*) F_{\beta^\top Y}(\beta^\top Z^*)\right], \\ &= (I) + (II) - 2(III) \quad \text{(say)}, \end{aligned}$$

where $Z^* \sim (1/2)P_X + (1/2)P_Y$. By the Fubini's theorem and the definition of Z^*, the first term (I) has the identity

$$\begin{aligned} (I) &= \mathbb{E}\left[\mathbb{1}(\beta^\top X_1 \le \beta^\top Z^*,\ \beta^\top X_2 \le \beta^\top Z^*)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\mathbb{1}(\beta^\top X_1 \le \beta^\top X_3, \beta^\top X_2 \le \beta^\top X_3)\right] + \frac{1}{2}\mathbb{E}\left[\mathbb{1}(\beta^\top X_1 \le \beta^\top Y_1, \beta^\top X_2 \le \beta^\top Y_1)\right]. \end{aligned}$$

Similarly,

$$\begin{aligned} (II) &= \mathbb{E}\left[\mathbb{1}(\beta^\top Y_1 \le \beta^\top Z^*, \beta^\top Y_2 \le \beta^\top Z^*)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\mathbb{1}(\beta^\top Y_1 \le \beta^\top Y_3, \beta^\top Y_2 \le \beta^\top Y_3)\right] + \frac{1}{2}\mathbb{E}\left[\mathbb{1}(\beta^\top Y_1 \le \beta^\top X_1, \beta^\top Y_2 \le \beta^\top X_1)\right] \end{aligned}$$

and

$$\begin{aligned} (III) &= \mathbb{E}\left[\mathbb{1}(\beta^\top X_1 \le \beta^\top Z^*, \beta^\top Y_1 \le \beta^\top Z^*)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\mathbb{1}(\beta^\top X_1 \le \beta^\top X_2, \beta^\top Y_1 \le \beta^\top X_2)\right] + \frac{1}{2}\mathbb{E}\left[\mathbb{1}(\beta^\top X_1 \le \beta^\top Y_2, \beta^\top Y_1 \le \beta^\top Y_2)\right]. \end{aligned}$$

We then apply Lemma 4.0.2 to obtain the desired result. □

Next we provide another expression for the CvM-statistic with a third-order kernel.

Lemma C.1.2 (Another expression for the CvM-statistic)**.** *Consider the kernel of order three*

$$h^\star_{\mathrm{CvM}}(x_1, x_2, x_3; y_1, y_2, y_3) \tag{C.9}$$

$$
\begin{aligned}
&= \frac{1}{2}\mathbb{E}\Big[\Big\{\mathbb{1}(\beta^\top x_1 \le \beta^\top x_3) - \mathbb{1}(\beta^\top y_1 \le \beta^\top x_3)\Big\} \\
&\qquad \times \Big\{\mathbb{1}(\beta^\top x_2 \le \beta^\top x_3) - \mathbb{1}(\beta^\top y_2 \le \beta^\top x_3)\Big\}\Big] \\
&\quad + \frac{1}{2}\mathbb{E}\Big[\Big\{\mathbb{1}(\beta^\top x_1 \le \beta^\top y_3) - \mathbb{1}(\beta^\top y_1 \le \beta^\top y_3)\Big\} \\
&\qquad \times \Big\{\mathbb{1}(\beta^\top x_2 \le \beta^\top y_3) - \mathbb{1}(\beta^\top y_2 \le \beta^\top y_3)\Big\}\Big].
\end{aligned}
$$

Let us define the corresponding U-statistic by

$$
U^\star_{\mathrm{CvM}} := \frac{1}{(m)_3(n)_3} \sum_{i_1,i_2,i_3=1}^{m,\neq} \sum_{j_1,j_2,j_3=1}^{n,\neq} h^\star_{\mathrm{CvM}}(X_{i_1}, X_{i_2}, X_{i_3}; Y_{j_1}, Y_{j_2}, Y_{j_3}).
$$

Then $U^\star_{\mathrm{CvM}}$ is an unbiased estimator of W_d^2. Furthermore when $\beta^\top X$ and $\beta^\top Y$ are continuous for λ-almost all $\beta \in \mathbb{S}^{d-1}$, it is simplified as

$$
U^\star_{\mathrm{CvM}} = \frac{1}{(m)_2(n)_2} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j_1,j_2=1}^{n,\neq} h_{\mathrm{CvM}}(X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2}). \tag{C.10}
$$

Proof. The unbiasedness property is trivial. We will show that (C.10) holds under the given conditions. Since there is no tie with probability one, we have

$$
\frac{1}{(m)_3} \sum_{i_1,i_2,i_3=1}^{m,\neq} \mathbb{E}_\beta[\mathbb{1}(\beta^\top X_{i_1} \le \beta^\top X_{i_3})\mathbb{1}(\beta^\top X_{i_2} \le \beta^\top X_{i_3})] = \frac{1}{3},
$$

$$
\frac{1}{(n)_3} \sum_{j_1,j_2,j_3=1}^{n,\neq} \mathbb{E}_\beta[\mathbb{1}(\beta^\top Y_{j_1} \le \beta^\top Y_{j_3})\mathbb{1}(\beta^\top Y_{j_2} \le \beta^\top Y_{j_3})] = \frac{1}{3}.
$$

Also the following identities are true

$$
\begin{aligned}
&\frac{2}{(m)_2 \cdot n} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j=1}^{n} \mathbb{E}_\beta[\mathbb{1}(\beta^\top X_{i_1} \le \beta^\top X_{i_2})\mathbb{1}(\beta^\top Y_j \le \beta^\top X_{i_2})] \\
&= 1 - \frac{1}{(m)_2 \cdot n} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j=1}^{n} \mathbb{E}_\beta[\mathbb{1}(\beta^\top X_{i_1} \le \beta^\top Y_j)\mathbb{1}(\beta^\top X_{i_2} \le \beta^\top Y_j)]
\end{aligned}
$$

and

$$
\frac{2}{m \cdot (n)_2} \sum_{i=1}^{m} \sum_{j_1,j_2=1}^{n,\neq} \mathbb{E}_\beta[\mathbb{1}(\beta^\top Y_{j_1} \le \beta^\top Y_{j_2})\mathbb{1}(\beta^\top X_i \le \beta^\top Y_{j_2})]
$$

$$= 1 - \frac{1}{m \cdot (n)_2} \sum_{i=1}^{m} \sum_{j_1, j_2 = 1}^{n, \neq} \mathbb{E}_\beta[\mathbb{1}(\beta^\top Y_{j_1} \leq \beta^\top X_i)\mathbb{1}(\beta^\top Y_{j_2} \leq \beta^\top X_i)].$$

After expanding the terms in $h^\star_{\mathrm{CvM}}$ and replacing the above identities, we can obtain

$$\begin{aligned} U^\star_{\mathrm{CvM}} = & \frac{1}{(m)_2 \cdot n} \sum_{i_1, i_2 = 1}^{m, \neq} \sum_{j=1}^{n} \mathbb{E}_\beta[\mathbb{1}(\beta^\top X_{i_1} \leq \beta^\top Y_j)\mathbb{1}(\beta^\top X_{i_2} \leq \beta^\top Y_j)] \\ & + \frac{1}{m \cdot (n)_2} \sum_{i=1}^{m} \sum_{j_1, j_2 = 1}^{n, \neq} \mathbb{E}_\beta[\mathbb{1}(\beta^\top Y_{j_1} \leq \beta^\top X_i)\mathbb{1}(\beta^\top Y_{j_2} \leq \beta^\top X_i)] - \frac{2}{3}, \\ = & \frac{1}{(m)_2 (n)_2} \sum_{i_1, i_2 = 1}^{m, \neq} \sum_{j_1, j_2 = 1}^{n, \neq} h_{\mathrm{CvM}}(X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2}). \end{aligned}$$

Hence the result follows. □

In the next lemma, we present an explicit expression for the variance of $U_{m,n}$, which will be used to bound the variance of the proposed statistic.

Lemma C.1.3 (Theorem 2 of Lee (1990) in Chapter 2)**.** *Let $U_{m,n}$ be a two-sample U-statistic based on a kernel having degrees k_1 and k_2. Then*

$$Var(U_{m,n}) = \sum_{c=0}^{k_1} \sum_{d=0}^{k_2} \frac{\binom{k_1}{c}\binom{k_2}{d}\binom{m-k_1}{k_1-c}\binom{n_2-k_2}{k_2-d}}{\binom{n_1}{k_1}\binom{n_2}{k_2}} \sigma^2_{c,d},$$

where $\sigma^2_{c,d}$ is defined similarly as (C.6).

Hoeffding (1952) identified a sufficient condition (indeed the necessary condition proved by Chung and Romano, 2013) under which the permutation distribution approximates the corresponding unconditional distribution. The condition is stated as follows:

Lemma C.1.4 (Theorem 5.1 of Chung and Romano (2013))**.** *Consider a sequence of random quantities X^n taking values in a sample space $\mathcal{M}^n$ and suppose that X^n has distribution P^n in $\mathcal{M}^n$. Let $\mathcal{G}_n$ be a finite group of transformation from $\mathcal{M}^n$ onto itself. Let $T_n = T_n(X^n)$ be any real valued statistic and ϖ_n be a random variable that is uniform on $\mathcal{G}_n$. Also, let ϖ'_n have the same distribution as ϖ_n, with X^n, ϖ_n and ϖ'_n mutually independent. Suppose, under P^n,*

$$(T_n(\varpi_n X^n), T_n(\varpi'_n X^n)) \xrightarrow{d} (T, T'), \tag{C.11}$$

where T and T' are independent, each with common cumulative distribution function $R(\cdot)$. Here, $\varpi_n X^n$ denotes the composition of X^n with ϖ_n and $\varpi'_n X^n$ is similarly defined. Let $\widehat{R}_n$ be the randomization

distribution function of T_n *defined by*

$$\widehat{R}_n(t) = \frac{1}{\#|\mathcal{G}_n|} \sum_{\varpi_n \in \mathcal{G}_n} \mathbb{1}\{T_n(\varpi_n X^n) \leq t\},$$

where $\#|\mathcal{G}_n|$ *denotes the cardinality of* $\mathcal{G}_n$*. Then, under* P^n*,*

$$\widehat{R}_n(t) \xrightarrow{p} R(t), \tag{C.12}$$

for every t *which is a continuity point of* $R(\cdot)$*. Conversely, if (C.12) holds for some limiting cumulative distribution function* $R(\cdot)$ *whenever* t *is a continuity point, then (C.11) holds.*

Chikkagoudar and Bhat (2014) studied the limiting distribution of a two-sample U-statistic under contiguous alternatives for the univariate case (see Theorem 3.1 therein and also Gregory, 1977). Here we extend their result to the multivariate case.

First we prepare for some notation. Let $P_{\theta_0}^N$ and $P_{\theta_0+bN^{-1/2}}^N$ denote the joint distribution of the pooled samples $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ under the null and contiguous alternative, respectively. Let $\lambda_{k,g}$ and $\phi_{k,g}(\cdot)$ be the eigenvalue and the corresponding eigenfunction satisfying the following integral equation

$$\mathbb{E}[g_{2,0}^*(x_1, X_2)\phi_{k,g}(X_2)] = \lambda_{k,g}\phi_{k,g}(x_1) \quad \text{for } k = 1, 2, \ldots,$$

where $g_{2,0}^*(\cdot,\cdot)$ is defined in (C.5) under the null hypothesis. For a sequence of random variables Z_N, we write $Z_N = o_{P_{\theta_0}^N}(1)$, if

$$\lim_{N\to\infty} P_{\theta_0}^N\big(|Z_N| \geq \epsilon\big) = 0,$$

for any $\epsilon > 0$. Then we have the following result.

Lemma C.1.5 (Limiting distribution under contiguous alternatives)**.** *Recall the two-sample* U*-statistic,* $U_{m,n}$*, given in (C.3). Consider the same assumptions used in Theorem C.1 and Theorem 4.4. Then under* $P_{\theta_0+bN^{-1/2}}^N$*,*

$$N(U_{m,n} - \mathbb{E}_{\theta_0}[U_{m,n}]) \xrightarrow{d} \frac{r(r-1)}{2\vartheta_X\vartheta_Y} \sum_{k=1}^{\infty} \lambda_{k,g}\{(\xi_k + \vartheta_X^{1/2} a_{k,g})^2 - 1\},$$

where

$$a_{k,g} = \int_{\mathbb{R}^d} \{b^\top 2\eta(x, \theta_0)\} p_{\theta_0}^{-1/2}(x)\phi_{k,g}(x) dP_{\theta_0}(x).$$

Proof. Let us denote the likelihood ratio by

$$L_{N,h} = \frac{\prod_{i=1}^m p_{\theta_0}(X_i) \prod_{j=1}^n p_{\theta_0+bN^{-1/2}}(Y_j)}{\prod_{i=1}^m p_{\theta_0}(X_i) \prod_{j=1}^n p_{\theta_0}(Y_j)} = \frac{\prod_{j=1}^n p_{\theta_0+bN^{-1/2}}(Y_j)}{\prod_{j=1}^n p_{\theta_0}(Y_j)}.$$

Then under the given conditions, one can establish

$$\log L_{N,h} = \frac{1}{\sqrt{n}} \sum_{i=1}^n h^\top \widetilde{\eta}(Y_i, \theta_0) - \frac{1}{2} h^\top I(\theta_0) h + o_{P_{\theta_0}^N}(1), \tag{C.13}$$

where $\widetilde{\eta}(x,\theta) = 2\eta(x,\theta)/p_\theta^{1/2}(x)$ (see Example 12.3.7 of Lehmann and Romano, 2006, for details). Then by Corollary 12.3.1 of Lehmann and Romano (2006), $P_{\theta_0}^N$ and $P_{\theta_0+bN^{-1/2}}^N$ are mutually contiguous.

Without loss of generality, we assume that $\mathbb{E}_{\theta_0}[U_{m,n}] = 0$ and denote the projection of $U_{m,n}$ under *condition 2* in Theorem C.1 by

$$\begin{aligned}\widehat{U}_{m,n} =& \frac{r(r-1)}{m(m-1)} \sum_{1\le i_1<i_2\le m} g_{2,0}^*(X_{i_1}, X_{i_2}) + \frac{r(r-1)}{n(n-1)} \sum_{1\le j_1<j_2\le m} g_{0,2}^*(Y_{j_1}, Y_{j_2}) \\ &+ \frac{r^2}{mn} \sum_{i=1}^m \sum_{j=1}^n g_{1,1}^*(X_i, Y_j).\end{aligned}$$

Then as in Lemma 2.2 of Chikkagoudar and Bhat (2014), it can be seen that

$$NU_{m,n} = N\widehat{U}_{m,n} + o_{P_{\theta_0}^N}(1),$$

and the same approximation holds under $P_{\theta_0+bN^{-1/2}}^N$ by contiguity. As a result, it is enough to study the limiting distribution of $N\widehat{U}_{m,n}$.

Now following the same steps in the proof of Theorem 3.1 in Chikkagoudar and Bhat (2014) and using (C.13), we can arrive at

$$N\widehat{U}_{m,n} \xrightarrow{d} \frac{r(r-1)}{2\vartheta_X\vartheta_Y} \sum_{k=1}^\infty \lambda_{k,g}\{(\xi_k + \vartheta_X^{1/2} a_{k,g})^2 - 1\},$$

under $P_{\theta_0+bN^{-1/2}}^N$. Hence the result follows. □

The next two lemmas are used for proving Theorem 4.8, which provides a lower bound for the minimax separation $\epsilon_{m,n}$. We begin by presenting a lower bound of the multivariate CvM-distance.

Lemma C.1.6 (Lower bound for the CvM-distance)**.** *The multivariate CvM-distance is lower bounded by*

$$W_d(P_X, P_Y) \ge \int_{\mathbb{S}^{d-1}} \left| \frac{1}{2} - \mathbb{P}\left(\beta^\top X \le \beta^\top Y\right) \right| d\lambda(\beta). \tag{C.14}$$

Proof. Let $\beta^\top Z$ have the distribution function $F_{\beta^\top X}(t)/2 + F_{\beta^\top Y}(t)/2$. First notice from the definition of the multivariate CvM-distance that

$$\begin{aligned} W_d^2 &= \mathbb{E}\Big[\Big\{F_{\beta^\top X}(\beta^\top Z) - F_{\beta^\top Y}(\beta^\top Z)\Big\}^2\Big] \\ &\geq \Big\{\mathbb{E}\Big[\Big|F_{\beta^\top X}(\beta^\top Z) - F_{\beta^\top Y}(\beta^\top Z)\Big|\Big]\Big\}^2, \end{aligned}$$

where the inequality follows by Jensen's inequality. Let us denote the expectation with respect to X_1, X_2, Y_1 (and X_1, Y_1, Y_2) by $\mathbb{E}_{X_1,X_2,Y_1}$ (and $\mathbb{E}_{X_1,Y_1,Y_2}$). Then from the definition of $\beta^\top Z$, we have

$$\begin{aligned} &\mathbb{E}\big[\big|F_{\beta^\top X}(\beta^\top Z) - F_{\beta^\top Y}(\beta^\top Z)\big|\big] \\ =\ &\frac{1}{2}\mathbb{E}\big[\big|F_{\beta^\top X}(\beta^\top X_1) - F_{\beta^\top Y}(\beta^\top X_1)\big|\big] + \frac{1}{2}\mathbb{E}\big[\big|F_{\beta^\top X}(\beta^\top Y_1) - F_{\beta^\top Y}(\beta^\top Y_1)\big|\big] \\ \geq\ &\frac{1}{2}\mathbb{E}_\beta\Big[\Big|\mathbb{E}_{X_1,X_2,Y_1}\Big\{\mathbb{1}(\beta^\top X_1 \leq \beta^\top X_2) - \mathbb{1}(\beta^\top Y_1 \leq \beta^\top X_2)\Big\}\Big|\Big] \\ &+ \frac{1}{2}\mathbb{E}_\beta\Big[\Big|\mathbb{E}_{X_1,Y_1,Y_2}\Big\{\mathbb{1}(\beta^\top X_1 \leq \beta^\top Y_2) - \mathbb{1}(\beta^\top Y_1 \leq \beta^\top Y_2)\Big\}\Big|\Big], \end{aligned}$$

where we used Jensen's inequality once again to obtain the lower bound. The last expression can be simplified based on the observation that $\mathbb{P}(\beta^\top X_1 \leq \beta^\top X_2) = \mathbb{P}(\beta^\top Y_1 \leq \beta^\top Y_2) = 1/2$ as

$$\mathbb{E}_\beta\Big[\Big|\frac{1}{2} - \mathbb{P}\left(\beta^\top X \leq \beta^\top Y\right)\Big|\Big].$$

Therefore,

$$W_d^2 \ \geq\ \Big\{\int_{\mathbb{S}^{d-1}} \Big|\frac{1}{2} - \mathbb{P}\left(\beta^\top X \leq \beta^\top Y\right)\Big| d\lambda(\beta)\Big\}^2,$$

which completes the proof. □

Consider two independent random vectors X^* and Y^* such that their first coordinates have normal distributions as $\xi_1 \sim N(\mu_{X^*}, 1)$ and $\xi_2 \sim N(\mu_{Y^*}, 1)$ and the other coordinates have the degenerate distribution at zero, i.e.

$$X^* := (\xi_1, 0, \ldots, 0)^\top \quad \text{and} \quad Y^* := (\xi_2, 0, \ldots, 0)^\top.$$

Given $\beta = (\beta_1, \ldots, \beta_d)^\top \in \mathbb{S}^{d-1}$, we have $\beta^\top X^* \sim N(\beta_1\mu_{X^*}, \beta_1^2)$ and $\beta^\top Y^* \sim N(\beta_1\mu_{Y^*}, \beta_1^2)$; therefore $\beta^\top X^*$ and $\beta^\top Y^*$ have continuous distributions for λ-almost all $\beta \in \mathbb{S}^{d-1}$. Under this setting, the multivariate CvM-distance is lower bounded as follows:

Lemma C.1.7 (Lower bound for the CvM-distance under a Gaussian model)**.** *Consider independent random vectors X^* and Y^* described above with $\mu_{X^*} = cm^{-1/2}$ and $\mu_{Y^*} = -cn^{-1/2}$ for some constant $c > 0$. Let us denote the corresponding distributions by P_{X^*} and P_{Y^*}. Then there exists another constant $C > 0$ independent of the dimension satisfying*

$$W_d(P_{X^*}, P_{Y^*}) \geq C\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right).$$

Furthermore, the lower bound is tight up to constant factors.

Proof. From Lemma C.1.6, it is enough to show

$$\int_{\mathbb{S}^{d-1}} \left|\frac{1}{2} - \mathbb{P}\left(\beta^\top X^* \leq \beta^\top Y^*\right)\right| d\lambda(\beta) \geq C\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right).$$

For any fixed $\beta \in \mathbb{S}^{d-1}$, we have $\beta^\top(X^* - Y^*) \sim N(\beta_1(\mu_{X^*} - \mu_{Y^*}), 2\beta_1^2)$. Let $\Phi(\cdot)$ and $\varphi(\cdot)$ denote the cumulative distribution function and the density function of the standard normal distribution respectively. Then

$$\begin{aligned}
\left|\frac{1}{2} - \mathbb{P}\left(\beta^\top X^* \leq \beta^\top Y^*\right)\right| &= \left|\frac{1}{2} - \Phi\left(-\operatorname{sign}(\beta_1) \cdot \frac{c}{\sqrt{2}}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right)\right)\right| \\
&\geq \frac{c}{\sqrt{2}}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) \cdot \varphi\left(\frac{c}{\sqrt{2}}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right)\right) \\
&\geq \frac{c}{\sqrt{2}}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right) \cdot \varphi\left(\frac{c}{2\sqrt{2}}\right),
\end{aligned}$$

This lower bound holds for λ-almost all $\beta \in \mathbb{S}^{d-1}$ and thus the result follows. To have an upper bound, notice that

$$\begin{aligned}
W_d^2(P_{X^*}, P_{Y^*}) \leq & \int_{\mathbb{S}^{d-1}} \sup_{t \in \mathbb{R}} \left(F_{\beta^\top X}(t) - F_{\beta^\top Y}(t)\right)^2 d\lambda(\beta) \\
\overset{(i)}{\leq} & \frac{1}{2}\int_{\mathbb{S}^{d-1}} \mathsf{KL}\big(N(\beta_1\mu_{X^*}, \beta_1^2), N(\beta_1\mu_{Y^*}, \beta_1^2)\big) d\lambda(\beta) \\
= & \frac{c^2}{2}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right)^2,
\end{aligned}$$

where $\mathsf{KL}(\cdot, \cdot)$ is the Kullback-Leibler divergence between two distributions and we used the Pinsker's inequality for (i) (e.g. Lemma 2.5 of Tsybakov, 2009). This shows the tightness of the lower bound. □

To prove Theorem 4.13, which presents a closed-form expression for $\tau^*_{p,q}$, we need to generalize Lemma 4.0.2 with three indicator functions as follows:

Lemma C.1.8 (Extension of Escanciano (2006)). *For arbitrary non-zero vectors $U_1, U_2, U_3 \in \mathbb{R}^d$, we have*

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} \prod_{i=1}^{3} \mathbb{1}(\beta^\top U_i \le 0) d\lambda(\beta) \\ = \ &\frac{1}{2} - \frac{1}{4\pi} \left[\mathsf{Ang}\left(U_1, U_2\right) + \mathsf{Ang}\left(U_1, U_3\right) + \mathsf{Ang}\left(U_2, U_3\right) \right]. \end{aligned}$$

Proof. The proof of this result can be found in Section C.4.3. □

C.4 Proofs

This section collects all proofs for the results in the main text. In addition to the notation given in the main text, we introduce further notation that will be used throughout this section.

Notation. We denote the probability measure under permutations by $\mathbb{P}_\varpi$. The expectation and variance with respect to $\mathbb{P}_\varpi$ are denoted by $\mathbb{E}_\varpi$ and Var_ϖ, respectively. We write the expectation with respect to the uniform probability measure λ on $\mathbb{S}^{d-1}$ by $\mathbb{E}_\beta$. The symbol $\#|A|$ stands for the cardinality of A. We denote the Kullback-Leibler divergence between two probability distributions P and Q by $\mathsf{KL}(P, Q)$. For $x, y \in \mathbb{R}$, we use $x \vee y$ and $x \wedge y$ to denote $\max\{x, y\}$ and $\min\{x, y\}$, respectively. Given a permutation ϖ of $\{1, \ldots, N\}$ and the pooled samples $\{Z_1, \ldots, Z_{m+n}\} = \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$, we may write $U_{\mathrm{CvM}}(Z_{\varpi(1)}, \ldots, Z_{\varpi(N)})$ or $U^{\varpi}_{\mathrm{CvM}}$ to denote the CvM-statistic computed based on $\mathcal{X}_m = \{Z_{\varpi(1)}, \ldots, Z_{\varpi(m)}\}$ and $\mathcal{Y}_n = \{Z_{\varpi(m+1)}, \ldots, Z_{\varpi(m+n)}\}$. For the original permutation, which is $\varpi = \{1, \ldots, N\}$, we write U_{CvM} or $U_{\mathrm{CvM}}(Z_1, \ldots, Z_1)$ to denote the CvM-statistic computed based on $\mathcal{X}_m = \{Z_1, \ldots, Z_m\}$ and $\mathcal{Y}_n = \{Z_1, \ldots, Z_{m+n}\}$. The similar notation will be used for other test statistics. In general, we will write $\widetilde{h}$ to denote the symmetrized version of a kernel h in the sense of (C.2). For any two real sequences $\{a_n\}$ and $\{b_n\}$, we write $b_n \gtrsim a_n$ or equivalently $a_n \lesssim b_n$ if there exists $C > 0$ such that $a_n \le C b_n$ for each n. $c, C, C_0, C_1, C_2, C_3, C_4, C_5$ are some universal constants whose values may differ in different places of this section.

C.4.1 Proof of Lemma 4.0.1

From the definition of W_d^2, it is clear to see that $W_d^2 \ge 0$ and it becomes zero if $P_X = P_Y$. For the other direction, we will show that if $W_d^2 = 0$, then X and Y have the same characteristic function:

$$\mathbb{E}_X\left[e^{it\beta^\top X}\right] = \mathbb{E}_Y\left[e^{it\beta^\top Y}\right] \quad \text{for all } (\beta, t) \in \mathbb{S}^{d-1} \times \mathbb{R},$$

which implies $P_X = P_Y$.

1. Univariate case

In the univariate case, $W^2 = 0$ implies that $F_X(t) = F_Y(t)$ for $d\{\vartheta_X F_X(t) + \vartheta_Y F_Y(t)\}$-almost all t, hence we conclude $P_X = P_Y$ (see also Lemma 4.1 of Lehmann, 1951).

2. Multivariate case

Recall that $\lambda(\cdot)$ is the uniform probability measure on $\mathbb{S}^{d-1}$. From the characteristic property of the univariate CvM-distance, $W_d^2 = 0$ implies that $\beta^\top X$ and $\beta^\top Y$ are identically distributed for λ-almost all $\beta \in \mathbb{S}^{d-1}$. Now, by continuity of the characteristic function, we conclude that

$$\mathbb{E}_X\left[e^{it\beta^\top X}\right] = \mathbb{E}_Y\left[e^{it\beta^\top Y}\right] \quad \text{for all } (\beta, t) \in \mathbb{S}^{d-1} \times \mathbb{R}.$$

C.4.2 Proof of Lemma 4.0.2

Here we provide an alternative proof of Lemma 4.0.2 based on the orthant probability for normal distribution. First we state a recent result on the bivariate normal distribution function presented by Monhor (2013).

Lemma C.1.9. *(Theorem 4 of Monhor, 2013) Let $(\xi_1, \xi_2)^\top$ has a bivariate normal distribution with expectation $(\mu_1, \mu_2)^\top = (0, 0)^\top$ and covariance matrix $[\sigma_{ij}]_{2\times 2}$ where $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = \rho$. Then for $0 < \rho < 1$ and $t > 0$,*

$$\mathbb{P}(\xi_1 \le t, \xi_2 \le t) \le \Phi^2(t) + \frac{1}{2\pi} \exp\left(-\frac{t^2}{1+\rho}\right) \arcsin(\rho) \quad \textit{and} \tag{C.15}$$

$$\mathbb{P}(\xi_1 \le t, \xi_2 \le t) \ge \Phi^2(t) + \frac{1}{2\pi} \exp\left(-t^2\right) \arcsin(\rho). \tag{C.16}$$

It is not difficult to see that a similar result can be obtained for $-1 < \rho \le 0$ as

$$\mathbb{P}(\xi_1 \le t, \xi_2 \le t) \le \Phi^2(t) - \frac{1}{2\pi} \exp\left(-\frac{t^2}{1+\rho}\right) \arcsin(-\rho) \quad \text{and} \tag{C.17}$$

$$\mathbb{P}(\xi_1 \le t, \xi_2 \le t) \ge \Phi^2(t) - \frac{1}{2\pi} \exp\left(-t^2\right) \arcsin(-\rho). \tag{C.18}$$

In fact, the inequalities (C.15), (C.16), (C.17) and (C.18) hold for any t. By taking $t \to 0$ in the previous inequalities, we have

$$\mathbb{P}(\xi_1 \le 0, \xi_2 \le 0) = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho) = \frac{1}{2} - \frac{1}{2\pi}\arccos(\rho), \tag{C.19}$$

for any $-1 \leq \rho \leq 1$. The above identity is classical and can be found in different places (e.g. Slepian, 1962; Childs, 1967; Xu et al., 2013).

Turning now to Lemma 4.0.2, let $\mathcal{Z}$ have a multivariate normal distribution with zero mean vector and identity covariance matrix. It is well-known that $\mathcal{Z}/\|\mathcal{Z}\|$ is uniformly distributed over $\mathbb{S}^{d-1}$ (e.g. page 15 of Anderson, 2003). This leads to the key observation that

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} \mathbb{1}(\beta^\top U_1 \leq 0)\mathbb{1}(\beta^\top U_2 \leq 0) d\lambda(\beta) \\ =\ & \mathbb{E}_{\mathcal{Z}}\left[\mathbb{1}(\mathcal{Z}^\top U_1 \leq 0)\mathbb{1}(\mathcal{Z}^\top U_2 \leq 0)\right], \end{aligned} \tag{C.20}$$

where $\mathbb{E}_{\mathcal{Z}}[\cdot]$ is the expectation with respect to $\mathcal{Z}$. Note that $(\mathcal{Z}^\top U_1, \mathcal{Z}^\top U_2)^\top$ follows a bivariate normal distribution with correlation matrix $[\varrho_{ij}]_{2\times 2}$ where $\varrho_{ij} = U_i^\top U_j/\{\|U_i\|\|U_j\|\}$. Using this connection and the equality (C.19), we can obtain the closed-form expression for the left-hand side of (C.20) and thus complete the proof.

C.4.3 Proof of Lemma C.1.8

To prove the results, we apply the same argument used in Section C.4.2. Let $\mathcal{Z}$ have a multivariate normal distribution with zero mean vector and identity covariance matrix. Then as in Section C.4.2,

$$\int_{\mathbb{S}^{d-1}} \prod_{i=1}^{3} \mathbb{1}(\beta^\top U_i \leq 0) d\lambda(\beta) = \mathbb{E}_{\mathcal{Z}}\left[\prod_{i=1}^{3} \mathbb{1}(\mathcal{Z}^\top U_i \leq 0)\right]. \tag{C.21}$$

Since $(\mathcal{Z}^\top U_1, \mathcal{Z}^\top U_2, \mathcal{Z}^\top U_3)^\top$ has a multivariate normal distribution with zero mean vector and correlation matrix $[\varrho_{ij}]_{3\times 3}$ with $\varrho_{ij} = U_i^\top U_j/\{\|U_i\|\|U_j\|\}$, the right-hand side of (C.21) can be computed based on orthant probabilities for normal distributions (e.g. Childs, 1967; Xu et al., 2013). This completes the proof.

C.4.4 Proof of Theorem 4.1

Since $\beta^\top X$ and $\beta^\top Y$ are assumed to have continuous distribution functions, $\beta^\top X_1, \beta^\top X_2$ and $\beta^\top X_3$ have distinct values with probability one. This is also true for $\beta^\top Y_1, \beta^\top Y_2$ and $\beta^\top Y_3$. Therefore, the following

identities hold for λ-almost all $\beta \in \mathbb{S}^{d-1}$.

$$\begin{aligned}
&\int \left(F_{\beta^\top X}(t)\right)^2 dF_{\beta^\top X}(t) = \mathbb{P}\left(\max\{\beta^\top X_1, \beta^\top X_2\} \le \beta^\top X_3\right) = \frac{1}{3},\\
&\int \left(F_{\beta^\top Y}(t)\right)^2 dF_{\beta^\top Y}(t) = \mathbb{P}\left(\max\{\beta^\top Y_1, \beta^\top Y_2\} \le \beta^\top Y_3\right) = \frac{1}{3},\\
&\int \left(F_{\beta^\top X}(t)\right)^2 dF_{\beta^\top Y}(t) = \mathbb{P}\left(\max\{\beta^\top X_1, \beta^\top X_2\} \le \beta^\top Y_1\right),\\
&\int \left(F_{\beta^\top Y}(t)\right)^2 dF_{\beta^\top X}(t) = \mathbb{P}\left(\max\{\beta^\top Y_1, \beta^\top Y_2\} \le \beta^\top X_1\right).
\end{aligned} \tag{C.22}$$

Also note that

$$\begin{aligned}
&\mathbb{P}\left(\max\{\beta^\top X_1, \beta^\top X_2\} \le \beta^\top Y_1\right) + \mathbb{P}\left(\max\{\beta^\top X_1, \beta^\top Y_1\} \le \beta^\top X_2\right)\\
&+ \mathbb{P}\left(\max\{\beta^\top X_2, \beta^\top Y_1\} \le \beta^\top X_1\right) = 1
\end{aligned}$$

and

$$\mathbb{P}\left(\max\{\beta^\top X_1, \beta^\top Y_1\} \le \beta^\top X_2\right) = \mathbb{P}\left(\max\{\beta^\top X_2, \beta^\top Y_1\} \le \beta^\top X_1\right).$$

These two identities give

$$\begin{aligned}
\int F_{\beta^\top X}(t) F_{\beta^\top Y}(t) dF_{\beta^\top X}(t) &= \mathbb{P}\left(\max\{\beta^\top X_1, \beta^\top Y_1\} \le \beta^\top X_2\right)\\
&= \frac{1}{2} - \frac{1}{2}\mathbb{P}\left(\max\{\beta^\top X_1, \beta^\top X_2\} \le \beta^\top Y_1\right).
\end{aligned} \tag{C.23}$$

Similarly,

$$\begin{aligned}
\int F_{\beta^\top X}(t) F_{\beta^\top Y}(t) dF_{\beta^\top Y}(t) &= \mathbb{P}\left(\max\{\beta^\top Y_1, \beta^\top X_1\} \le \beta^\top Y_2\right)\\
&= \frac{1}{2} - \frac{1}{2}\mathbb{P}\left(\max\{\beta^\top Y_1, \beta^\top Y_2\} \le \beta^\top X_1\right).
\end{aligned} \tag{C.24}$$

Now, combine (C.22), (C.23) and (C.24) to have

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(F_{\beta^\top X}(t) - F_{\beta^\top Y}(t)\right)^2 d\{\vartheta_X F_{\beta^\top X}(t) + \vartheta_Y F_{\beta^\top Y}(t)\} d\lambda(\beta)\\
= &\int_{\mathbb{S}^{d-1}} \mathbb{P}\left(\max\{\beta^\top X_1, \beta^\top X_2\} \le \beta^\top Y_1\right) d\lambda(\beta)\\
&+ \int_{\mathbb{S}^{d-1}} \mathbb{P}\left(\max\{\beta^\top Y_1, \beta^\top Y_2\} \le \beta^\top X_1\right) d\lambda(\beta) - \frac{2}{3}.
\end{aligned}$$

Hence,

$$W_d^2 = \mathbb{E}\left[\mathbb{1}(\beta^\top X_1 \le \beta^\top Y_1, \beta^\top X_2 \le \beta^\top Y_1)\right] + \mathbb{E}\left[\mathbb{1}(\beta^\top Y_1 \le \beta^\top X_1, \beta^\top Y_2 \le \beta^\top X_1)\right] - \frac{2}{3}.$$

Then apply Lemma 4.0.2 to obtain the result.

C.4.5 Proof of Theorem 4.2

We first show that h is degenerate under H_0. Then apply the limit theorem for two-sample degenerate U-statistics (Bhat, 1995).

1. Degeneracy

Recall the definition of the kernel h_{CvM}, i.e.

$$h_{\mathrm{CvM}}(x_1, x_2; y_1, y_2) = \frac{1}{3} - \frac{1}{2\pi}\mathsf{Ang}(x_1 - y_1, x_2 - y_1) - \frac{1}{2\pi}\mathsf{Ang}(y_1 - x_1, y_2 - x_1).$$

We also recall the symmetrized version of h_{CvM} by $\widetilde{h}_{\mathrm{CvM}}$ in the sense of (C.2), i.e.

$$\widetilde{h}_{\mathrm{CvM}}(x_1, x_2; y_1, y_2) = \frac{1}{2}h_{\mathrm{CvM}}(x_1, x_2; y_1, y_2) + \frac{1}{2}h_{\mathrm{CvM}}(x_2, x_1; y_2, y_1).$$

We first focus on the univariate case where $x_1, x_2, y_1, y_2 \in \mathbb{R}$ and make a connection to Lehmann's two-sample statistic (Lehmann, 1951). Let $\widetilde{h}^{(1)}_{\mathrm{CvM}}$ denote the symmetrized h_{CvM} for the univariate case, that can be written as

$$\widetilde{h}^{(1)}_{\mathrm{CvM}}(x_1, x_2; y_1, y_2) := \frac{1}{2}\Big\{\mathbb{1}(\max\{x_1, x_2\} \le y_1) + \mathbb{1}(\max\{x_1, x_2\} \le y_2) + \mathbb{1}(\max\{y_1, y_2\} \le x_1) + \mathbb{1}(\max\{y_1, y_2\} \le x_2)\Big\} - \frac{2}{3}.$$

From the following identity,

$$\begin{aligned} & \mathbb{1}(\max\{x_1, x_2\} \le \min\{y_1, y_2\}) + \mathbb{1}(\max\{y_1, y_2\} \le \min\{x_1, x_2\}) \\ = \; & \mathbb{1}(\max\{x_1, x_2\} \le y_1) + \mathbb{1}(\max\{x_1, x_2\} \le y_2) \\ + \; & \mathbb{1}(\max\{y_1, y_2\} \le x_1) + \mathbb{1}(\max\{y_1, y_2\} \le x_2) - 1, \end{aligned}$$

the univariate kernel has another expression as

$$\begin{aligned} 2\widetilde{h}^{(1)}_{\mathrm{CvM}}(x_1, x_2; y_1, y_2) &= \mathbb{1}(\max\{x_1, x_2\} \le \min\{y_1, y_2\}) \\ &\quad + \mathbb{1}(\max\{y_1, y_2\} \le \min\{x_1, x_2\}) - \frac{1}{3}. \end{aligned}$$

Thus $\widetilde{h}^{(1)}_{\mathrm{CvM}}$ is equivalent to the kernel for Lehmann's two-sample statistic (Lehmann, 1951). Using this connection and the known results for Lehmann's two-sample statistic, we have

$$\begin{aligned} \widetilde{h}^{(1)}_{\mathrm{CvM},1,0}(x_1) &:= \mathbb{E}\left[\widetilde{h}^{(1)}_{\mathrm{CvM}}(x_1, X_2; Y_1, Y_2)\right] = 0, \\ \widetilde{h}^{(1)}_{\mathrm{CvM},0,1}(y_1) &:= \mathbb{E}\left[\widetilde{h}^{(1)}_{\mathrm{CvM}}(X_1, X_2; y_1, Y_2)\right] = 0, \end{aligned} \tag{C.25}$$

for any $x_1, y_1 \in \mathbb{R}$ under H_0. See Chapter 4 of Bhat (1995) for details.

Let us now turn to multivariate cases where $x_1, x_2, y_1, y_2 \in \mathbb{R}^d$. By the definition of $\widetilde{h}_{\mathrm{CvM}}$, we have

$$\widetilde{h}_{\mathrm{CvM}}(x_1, x_2, y_1, y_2) = \int_{\mathbb{S}^{d-1}} \widetilde{h}^{(1)}_{\mathrm{CvM}}(\beta^\top x_1, \beta^\top x_2; \beta^\top y_1, \beta^\top x_2) d\lambda(\beta).$$

Now the Fubini's theorem combined with (C.25) gives

$$\begin{aligned} &\mathbb{E}\left[\widetilde{h}^{(1)}_{\mathrm{CvM}}(\beta^\top x_1, \beta^\top X_2; \beta^\top Y_1, \beta^\top Y_2)\right] \\ =\ &\mathbb{E}\left[\widetilde{h}^{(1)}_{\mathrm{CvM}}(\beta^\top X_1, \beta^\top X_2; \beta^\top y_1, \beta^\top Y_2)\right] = 0, \end{aligned}$$

for λ-almost all $\beta \in \mathbb{S}^{d-1}$. As a consequence, it is seen that

$$\begin{aligned} \widetilde{h}_{\mathrm{CvM},1,0}(x_1) &:= \mathbb{E}\left[\widetilde{h}_{\mathrm{CvM}}(x_1, X_2; Y_1, Y_2)\right] \\ &= \int_{\mathbb{S}^{d-1}} \mathbb{E}\left[\widetilde{h}^{(1)}_{\mathrm{CvM}}(\beta^\top x_1, \beta^\top X_2; \beta^\top Y_1, \beta^\top Y_2)\right] d\lambda(\beta) = 0, \\ \widetilde{h}_{\mathrm{CvM},0,1}(y_1) &:= \mathbb{E}\left[\widetilde{h}_{\mathrm{CvM}}(X_1, X_2; y_1, Y_2)\right] \\ &= \int_{\mathbb{S}^{d-1}} \mathbb{E}\left[\widetilde{h}^{(1)}_{\mathrm{CvM}}(\beta^\top X_1, \beta^\top X_2; \beta^\top y_1, \beta^\top Y_2)\right] d\lambda(\beta) = 0. \end{aligned}$$

On the other hand,

$$\begin{aligned} \widetilde{h}_{\mathrm{CvM},2,0}(x_1, x_2) &:= \mathbb{E}\left[\widetilde{h}_{\mathrm{CvM}}(x_1, x_2; Y_1, Y_2)\right] \\ &= \frac{1}{2}\int_{\mathbb{S}^{d-1}} \left(1 - F_{\beta^\top X}(\max\{\beta^\top x_1, \beta^\top x_2\})\right)^2 d\lambda(\beta) \end{aligned}$$

$$+\frac{1}{2}\int_{\mathbb{S}^{d-1}} F^2_{\beta^\top X}(\min\{\beta^\top x_1, \beta^\top x_2\})d\lambda(\beta) - \frac{1}{6},$$

$$\begin{aligned}
\widetilde{h}_{\mathrm{CvM},0,2}(y_1, y_2) &:= \mathbb{E}\left[\widetilde{h}_{\mathrm{CvM}}(X_1, X_2; y_1, y_2)\right], \\
&= \frac{1}{2}\int_{\mathbb{S}^{d-1}} \left(1 - F_{\beta^\top Y}(\max\{\beta^\top y_1, \beta^\top y_2\})\right)^2 d\lambda(\beta) \\
&\quad + \frac{1}{2}\int_{\mathbb{S}^{d-1}} F^2_{\beta^\top Y}(\min\{\beta^\top y_1, \beta^\top y_2\})d\lambda(\beta) - \frac{1}{6}, \\
\widetilde{h}_{\mathrm{CvM},1,1}(x_1, y_1) &:= \mathbb{E}\left[\widetilde{h}_{\mathrm{CvM}}(x_1, X_2; y_1, Y_2)\right] \\
&= -\frac{1}{2}\widetilde{h}_{\mathrm{CvM},2,0}(x_1, y_1).
\end{aligned}$$

Note that $\widetilde{h}_{\mathrm{CvM},2,0}(x_1, x_2) \neq 0$ for some (x_1, x_2). For example, when $x_1 = x_2$, it is seen that

$$\frac{1}{2}\left\{1 - F_{\beta^\top X}(\beta^\top x_1)\right\}^2 + \frac{1}{2}F^2_{\beta^\top X}(\beta^\top x_1) - \frac{1}{6} \;\geq\; \frac{1}{12} \quad \text{for all } \beta \in \mathbb{S}^{d-1},$$

which implies $\widetilde{h}_{\mathrm{CvM},2,0}(x_1, x_1) \geq 1/12$. By the continuity of $\widetilde{h}_{\mathrm{CvM},2,0}$ at (x_1, x_1), there exist a set with nonzero measure such that $\widetilde{h}_{\mathrm{CvM},2,0}(x_1, x_2) > 0$. Therefore, we conclude that $\widetilde{h}_{\mathrm{CvM}}$ (and h_{CvM}) has degeneracy of order one under H_0.

2. Limiting distribution of the U-statistic

To obtain the limiting null distribution of U_{CvM}, we apply the result given in Chapter 3 of Bhat (1995) to have

$$NU_{\mathrm{CvM}} \xrightarrow{d} \frac{1}{\vartheta_X}\sum_{k=1}^{\infty}\lambda_k(\xi_k^2 - 1) + \frac{1}{\vartheta_Y}\sum_{k=1}^{\infty}\lambda_k(\xi_k'^2 - 1) - \frac{2}{\sqrt{\vartheta_X\vartheta_Y}}\sum_{k=1}^{\infty}\lambda_k\xi_k\xi_k',$$

where $\xi_k, \xi_k' \overset{i.i.d.}{\sim} N(0,1)$. Based on the observation that

$$\sqrt{\vartheta_Y}\xi_k - \sqrt{\vartheta_X}\xi_k' \sim N(0,1),$$

the result follows.

Remark C.1. *The eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ may depend on the underlying distribution, which implies that the test statistic is not distribution-free even asymptotically. Nevertheless, for the univariate continuous case, explicit expressions for the eigenvalues and the eigenfunctions are available as $\lambda_i = 2/(i\pi)^2$ and $\phi_i(x) = \sqrt{2}\cos(i\pi x)$ for $i = 1, 2, \ldots$ (e.g. Chikkagoudar and Bhat, 2014).*

C.4.6 Proof of Theorem 4.3

Let us write

$$\widetilde{h}_{\mathrm{CvM},1,0}(x) = \mathbb{E}[\widetilde{h}_{\mathrm{CvM}}(x, X_1; Y_1, Y_2)] \quad \text{and}$$
$$\widetilde{h}_{\mathrm{CvM},0,1}(y) = \mathbb{E}[\widetilde{h}_{\mathrm{CvM}}(X_1, X_2; y, Y_1)].$$

By Hoeffding's decomposition of a two-sample U-statistic (e.g. page 40 of Lee, 1990), the CvM-statistic can be approximated by

$$U_{\mathrm{CvM}} - W_d^2 = \frac{2}{m}\sum_{i=1}^{m} \widetilde{h}_{\mathrm{CvM},1,0}(X_i) + \frac{2}{n}\sum_{j=1}^{n} \widetilde{h}_{\mathrm{CvM},0,1}(Y_j) + O_{\mathbb{P}}(N^{-1}).$$

Then the result follows by the central limit theorem.

C.4.7 Proof of Theorem 4.4

This is a direct consequence of Lemma C.1.5 with $r = 2$, which completes the proof.

C.4.8 Proof of Theorem 4.5

Under the null hypothesis, we need to verify the conditions given in Theorem C.1 of this supplementary document. Indeed, these conditions are satisfied with $r = 2$ as in the proof of Theorem 4.2. Therefore the permutation distribution of the U-statistic converges in probability to the limiting null distribution of the unconditional statistic, which is $P_{\mathrm{CvM}}^{(\mathrm{single})} = P_{\mathrm{CvM}}^{(\mathrm{mix})}$ where we recall that

- $P_{\mathrm{CvM}}^{(\mathrm{single})}$: the limiting null distribution of NU_{CvM} based on i.i.d. samples from the single distribution P_X.
- $P_{\mathrm{CvM}}^{(\mathrm{mix})}$: the limiting null distribution of NU_{CvM} based on i.i.d. samples from the mixture distribution $\vartheta_X P_X + \vartheta_Y P_Y$.

By construction, the oracle test statistic has the same limit under H_0 and the result follows in this case.

Next we focus on the alternative hypothesis. In the next lemma, we first show that condition (C.8) is satisfied for the CvM-statistic, meaning that the difference between the two CvM-statistics — one is based on the randomly permuted original samples and the other is based on the corresponding coupled *i.i.d.* samples — is asymptotically negligible.

Lemma C.1.10 (Coupling for the CvM-statistic)**.** *Consider the two sets of samples* $\{Z_1, \ldots, Z_N\}$ *and* $\{\overline{Z}_{\varpi_0(1)}, \ldots, \overline{Z}_{\varpi_0(N)}\}$ *from Algorithm 4 where* $P_X := P_{X,N}$ *and* $P_Y := P_{Y,N}$ *can change arbitrary with* N.

Denote their random permutations by $\{Z_{\varpi(1)}, \ldots, Z_{\varpi(N)}\}$ *and* $\{\overline{Z}_{\varpi(\varpi_0(1))}, \ldots, \overline{Z}_{\varpi(\varpi_0(N))}\}$. *Then, under the assumption that* $m/N - \vartheta_X = O(N^{-1/2})$, *we have*

$$\begin{aligned} & NU_{\mathrm{CvM}}(Z_{\varpi(1)}, \ldots, Z_{\varpi(N)}) - NU_{\mathrm{CvM}}(\overline{Z}_{\varpi(\varpi_0(1))}, \ldots, \overline{Z}_{\varpi(\varpi_0(N))}) \\ & \xrightarrow{p} 0. \end{aligned} \tag{C.26}$$

The proof of the above lemma can be found in Section C.4.9 of this material. It is important to note that the convergence result (C.26) holds for any sequence of underlying distributions $P_{X,N}$ and $P_{Y,N}$, which should be clear from the proof of Lemma C.1.10. This means that the convergence result (C.26) is true under both fixed and contiguous alternatives. For simplicity, let us write

$$\begin{aligned} U_{N,\varpi} &= NU_{\mathrm{CvM}}(Z_{\varpi(1)}, \ldots, Z_{\varpi(N)}), \\ \overline{U}_{N,\varpi} &= NU_{\mathrm{CvM}}(\overline{Z}_{\varpi(\varpi_0(1))}, \ldots, \overline{Z}_{\varpi(\varpi_0(N))}). \end{aligned} \tag{C.27}$$

By letting ϖ' be an i.i.d. copy of ϖ, we also denote by $U_{N,\varpi'}$ and $\overline{U}_{N,\varpi'}$ the test statistics computed similarly as $U_{N,\varpi}$ and $\overline{U}_{N,\varpi}$, respectively, but replacing ϖ with ϖ'. Now by invoking (i) Theorem 4.2, (ii) Theorem C.1 and (iii) Hoeffding's condition in Lemma C.1.4, we know that under the fixed alternative

$$(\overline{U}_{N,\varpi}, \overline{U}_{N,\varpi'}) \xrightarrow{d} (U, U'),$$

where U, U' are i.i.d. random variables from the Gaussian chaos distribution $P_{\mathrm{CvM}}^{(\mathrm{mix})}$. By Slutsky's theorem, the convergence result (C.26) further implies that

$$(U_{N,\varpi}, U_{N,\varpi'}) \xrightarrow{d} (U, U').$$

By applying Hoeffding's condition in Lemma C.1.4 again, we know that the permutation distribution of $U_{N,\varpi}$ converges to $P_{\mathrm{CvM}}^{(\mathrm{mix})}$ in probability. As in the main text, let $q_{\alpha,\mathrm{CvM}}^{(\mathrm{mix})}$ be the upper $1-\alpha$ quantile of $P_{\mathrm{CvM}}^{(\mathrm{mix})}$. Then it follows that $c_{\alpha,\mathrm{CvM}}^{*} \xrightarrow{p} q_{\alpha,\mathrm{CvM}}^{(\mathrm{mix})}$ (see, e.g. Lemma 11.2.1 of Lehmann and Romano, 2006) and also $c_{\alpha,\mathrm{CvM}} \xrightarrow{p} q_{\alpha,\mathrm{CvM}}^{(\mathrm{mix})}$. Therefore the result follows under the fixed alternative.

Now it remains to prove the same statement under the contiguous alternative where $P_X = P_{\theta_0}$ and $P_Y = P_{\theta_0 + bN^{-1/2}}$. Using the same argument as before, under the null where $P_X = P_Y = P_{\theta_0}$, it can be seen that the permutation distribution of $U_{N,\varpi}$ converges to the Gaussian chaos distribution $P_{\mathrm{CvM}}^{(\mathrm{single})}$ in probability. By the property of contiguity (e.g. Theorem 12.3.2 of Lehmann and Romano, 2006), the limiting permutation distribution of $U_{N,\varpi}$ does not change under the contiguous alternative. Based on this result,

observe that the reverse direction of Hoeffding's condition in Lemma C.1.4 gives

$$(U_{N,\varpi}, U_{N,\varpi'}) \xrightarrow{d} (U, U'),$$

under the contiguous alternative where U, U' are i.i.d. random variables from $P_{\mathrm{CvM}}^{(\mathrm{single})}$. Then, under the same setting, Slutsky's theorem together with Lemma C.1.10 yields

$$(\overline{U}_{N,\varpi}, \overline{U}_{N,\varpi'}) \xrightarrow{d} (U, U').$$

In other words, the unconditional distribution of $\overline{U}_{N,\varpi}$ has the same Gaussian chaos limit as the permutation distribution of $U_{N,\varpi}$ under the contiguous alternative, that is to say $P_{\mathrm{CvM}}^{(\mathrm{mix})} = P_{\mathrm{CvM}}^{(\mathrm{single})}$. This completes the proof of Theorem 4.5.

C.4.9 Proof of Lemma C.1.10

In this section, we prove Lemma C.1.10 that is key to the proof of Theorem 4.5. Using the result in Lemma C.1.2, we work with the third-order kernel $h_{\mathrm{CvM}}^{\star}$ in (C.9). First notice that the expectations of both $U_{\mathrm{CvM}}(Z_{\varpi(1)}, \ldots, Z_{\varpi(N)})$ and $U_{\mathrm{CvM}}(\overline{Z}_{\varpi(\varpi_0(1))}, \ldots, \overline{Z}_{\varpi(\varpi_0(N))})$ are zero. To see this, putting $\mathcal{E} = \{\beta, Z_1, \ldots, Z_N, \varpi(2), \varpi(3), \varpi(m+2)\}$, let us consider the conditional expectation given by

$$\begin{aligned} f(\mathcal{E}) = \; & \mathbb{E}_{\varpi(1),\varpi(m+1)}\big[\big\{\mathbb{1}(\beta^\top Z_{\varpi(1)} \le \beta^\top Z_{\varpi(3)}) \\ & - \mathbb{1}(\beta^\top Z_{\varpi(m+1)} \le \beta^\top Z_{\varpi(3)})\big\} \big| \; \mathcal{E}\big]. \end{aligned}$$

By the linearity of expectation, it is clear to see that $f(\mathcal{E})$ is zero for any $\mathcal{E}$. As a result, the law of total expectation gives

$$\begin{aligned} & \mathbb{E}\big[\{\mathbb{1}(\beta^\top Z_{\varpi(1)} \le \beta^\top Z_{\varpi(3)}) - \mathbb{1}(\beta^\top Z_{\varpi(m+1)} \le \beta^\top Z_{\varpi(3)})\} \\ & \qquad \times \{\mathbb{1}(\beta^\top Z_{\varpi(2)} \le \beta^\top Z_{\varpi(3)}) - \mathbb{1}(\beta^\top Z_{\varpi(m+2)} \le \beta^\top Z_{\varpi(3)})\}\big] \\ = \; & \mathbb{E}\big[f(\mathcal{E}) \times \{\mathbb{1}(\beta^\top Z_{\varpi(2)} \le \beta^\top Z_{\varpi(3)}) - \mathbb{1}(\beta^\top Z_{\varpi(m+2)} \le \beta^\top Z_{\varpi(3)})\}\big] = 0. \end{aligned}$$

By applying the same logic to the other terms, it can be seen that the expectations of both test statistics are zero.

Let us recall the notation $U_{N,\varpi}$ and $\overline{U}_{N,\varpi}$ from (C.27). Based on the previous observation, it suffices to prove that the expected value of the squared difference between $U_{N,\varpi}$ and $\overline{U}_{N,\varpi}$ converges to zero as

$$\mathbb{E}\big[(U_{N,\varpi} - \overline{U}_{N,\varpi})^2\big] = o(1). \tag{C.28}$$

If this is the case, then Chebyshev's inequality guarantees the convergence result (C.26) and completes the proof.

For simplicity, write

$$\begin{aligned}
& d^{\star}_{\varpi}(i_1, i_2, i_3; j_1, j_2, j_3) \\
= \; & h^{\star}_{\mathrm{CvM}}\Big(Z_{\varpi(i_1)}, Z_{\varpi(i_2)}, Z_{\varpi(i_3)}; Z_{\varpi(j_1+m)}, Z_{\varpi(j_2+m)}, Z_{\varpi(j_3+m)}\Big) \\
& - h^{\star}_{\mathrm{CvM}}\Big(\overline{Z}_{\varpi(\varpi_0(i_1))}, \overline{Z}_{\varpi(\varpi_0(i_2))}, \overline{Z}_{\varpi(\varpi_0(i_3))}; \\
& \qquad \overline{Z}_{\varpi(\varpi_0(j_1+m))}, \overline{Z}_{\varpi(\varpi_0(j_2+m))}, \overline{Z}_{\varpi(\varpi_0(j_3+m))}\Big).
\end{aligned}$$

Then the squared difference can be written as

$$\begin{aligned}
(U_{N,\varpi} - \overline{U}_{N,\varpi})^2 \; = \; & \frac{N^2}{(m)_3^2 (n)_3^2} \times \\
& \sum_{i_1,i_2,i_3=1}^{m,\neq} \sum_{j_1,j_2,j_3=1}^{n,\neq} \sum_{i'_1,i'_2,i'_3=1}^{m,\neq} \sum_{j'_1,j'_2,j'_3=1}^{n,\neq} d^{\star}_{\varpi}(i_1, i_2, i_3; j_1, j_2, j_3) d^{\star}_{\varpi}(i'_1, i'_2, i'_3; j'_1, j'_2, j'_3).
\end{aligned}$$

Further write

$$\mathcal{I}_3 = \{i_1, i_2, i_3\} \cap \{i'_1, i'_2, i'_3\} \quad \text{and} \quad \mathcal{J}_3 = \{j_1, j_2, j_3\} \cap \{j'_1, j'_2, j'_3\}. \tag{C.29}$$

We analyze the expected value of the summand depending on the cardinality of $\mathcal{I}_3$ and $\mathcal{J}_3$. First consider the cases where $\#|\mathcal{I}_3| + \#|\mathcal{J}_3| \leq 1$. By the law of total expectation and putting $\mathcal{E}' = \{\beta, Z_1, \ldots, Z_N, \overline{Z}_1, \ldots, \overline{Z}_N\}$, it can be shown that

$$\mathbb{E}\left[d^{\star}_{\varpi}(i_1, i_2, i_3; j_1, j_2, j_3) d^{\star}_{\varpi}(i'_1, i'_2, i'_3; j'_1, j'_2, j'_3) | \mathcal{E}'\right] = 0,$$

Thus the unconditional expectation is also zero in these cases.

Next consider the cases where $\#|\mathcal{I}_3| + \#|\mathcal{J}_3| = 2$. More specifically, we split the cases into

- $\mathcal{C}_a = \{i_1, \ldots, i'_3, j_1, \ldots, j'_3 : \#|\mathcal{I}_3| = 2 \text{ and } \#|\mathcal{J}_3| = 0\}$,

- $\mathcal{C}_b = \{i_1, \ldots, i_3', j_1, \ldots, j_3' : \#|\mathcal{I}_3| = 0 \text{ and } \#|\mathcal{J}_3| = 2\}$,

- $\mathcal{C}_c = \{i_1, \ldots, i_3', j_1, \ldots, j_3' : \#|\mathcal{I}_3| = 1 \text{ and } \#|\mathcal{J}_3| = 1\}$.

Suppose there are B_1 different observations between

$$\{Z_{\varpi(1)}, \ldots, Z_{\varpi(m)}\} \quad \text{and} \quad \{\overline{Z}_{\varpi(\varpi_0(1))}, \ldots, \overline{Z}_{\varpi(\varpi_0(m))}\}$$

and B_2 different observations between

$$\{Z_{\varpi(m+1)}, \ldots, Z_{\varpi(m+n)}\} \quad \text{and} \quad \{\overline{Z}_{\varpi(\varpi_0(m+1))}, \ldots, \overline{Z}_{\varpi(\varpi_0(m+n))}\}.$$

Hence, we have $D = B_1 + B_2$ different observations in total between the original samples and the coupled samples. In these cases, it can be seen that

$$\begin{aligned}
\#|\mathcal{C}_a| &\lesssim B_1 m^3 n^6 + B_2 m^4 n^5, \\
\#|\mathcal{C}_b| &\lesssim B_1 m^5 n^4 + B_2 m^6 n^3, \\
\#|\mathcal{C}_c| &\lesssim B_1 m^4 n^5 + B_2 m^5 n^4.
\end{aligned}$$

Also note that the number of the other cases such that $\#|\mathcal{I}_3| + \#|\mathcal{J}_3| > 2$ are at most $O(N^9)$. Combining the previous observations together with the boundedness of the kernel $d_{\varpi}^{\star}$ yields

$$\mathbb{E}\big[(U_{N,\varpi} - \overline{U}_{N,\varpi})^2\big] \lesssim \mathbb{E}[D/N].$$

On the other hand, under the assumption that $m/N - \vartheta_X = O(N^{-1/2})$, we have $\mathbb{E}[D] = O(\sqrt{N})$ (e.g. Chung and Romano, 2013), which in turn gives

$$\mathbb{E}\big[(U_{N,\varpi} - \overline{U}_{N,\varpi})^2\big] = O\left(\frac{1}{\sqrt{N}}\right) = o(1).$$

This completes the proof of Lemma C.1.10.

C.4.10 Proof of Theorem 4.6

The type I error control of the oracle test and the permutation test are obvious and well-known (Chapter 15 of Lehmann and Romano, 2006). Hence we focus on the asymptotic power of the tests. When P_X and P_Y are fixed, it is not difficult to show that all the tests have asymptotic power equal to one; hence the result

follows. In fact, we can prove a stronger result that even if the CvM-distance between P_X and P_Y shrinks to zero as the sample size increases, the given tests can be consistent (see, e.g., Theorem 5.5).

Next turning to the contiguous alternative where $P_X = P_{\theta_0}$ and $P_Y = P_{\theta_0 + bN^{-1/2}}$, let us recall that $q^{(\text{single})}_{\alpha,\text{CvM}}$ and $q^{(\text{mix})}_{\alpha,\text{CvM}}$ are the $1-\alpha$ quantiles of $P^{(\text{single})}_{\text{CvM}}$ or $P^{(\text{mix})}_{\text{CvM}}$, respectively. Then as we proved in Theorem 4.5, we have that

$$c_{\alpha,\text{CvM}} \xrightarrow{p} q^{(\text{mix})}_{\alpha,\text{CvM}} = q^{(\text{single})}_{\alpha,\text{CvM}} \quad \text{and} \quad c^*_{\alpha,\text{CvM}} \xrightarrow{p} q^{(\text{mix})}_{\alpha,\text{CvM}} = q^{(\text{single})}_{\alpha,\text{CvM}},$$

under the contiguous alternative. Then the result follows by Theorem 4.4 and Slutsky's theorem.

C.4.11 Proof of Theorem 4.7

The outline of the proof of Theorem 4.7 is as follows. We start by presenting two lemmas where we bound the variance of U_{CvM} (Lemma C.1.11) and study the two moments of U_{CvM} under permutations (Lemma C.1.12). Based on these two lemmas, we first prove the statement on the CvM test, i.e. $\lim_{m,n\to\infty} \inf_{G_N} \mathbb{E}_1 [\phi_{\text{CvM}}] = 1$. We then turn to the energy test and show that $\lim_{m,n\to\infty} \inf_{G_N} \mathbb{E}_1 [\phi_{\text{Energy}}] \leq \alpha$. Before we start, we present one remark.

Remark C.2. *From the integral representations in (4.3) and (4.4) of the main text, it is seen that* $E_d(P_{X,N}, P_{Y,N}) = (1-\epsilon)E_d(Q_X, Q_Y)$ *and* $W_d(P_{X,N}, P_{Y,N}) \geq (1-\epsilon)W_d(Q_X, Q_Y)$*, which are positive provided that* $Q_X \neq Q_Y$*. This explains that the poor performance of the energy test is not because of lack of signal in the contamination model but because of non-robustness of the energy test statistic.*

Lemma C.1.11 (Variance of U_{CvM})**.** *Consider the CvM-statistic in (4.7). Then there exist universal constants* $C_1, C_2, C_3, C_4 > 0$ *such that*

$$Var[U_{\text{CvM}}] \;\leq\; C_1 \mathbb{E}[U_{\text{CvM}}] \left(\frac{1}{m} + \frac{1}{n}\right) + \frac{C_2}{m^2} + \frac{C_3}{n^2} + \frac{C_4}{mn}.$$

Proof. For this proof, it is more convenient to work with the third-order kernel given in (C.9). Let $\widetilde{h}^\star_{\text{CvM}}$ be the symmetrized kernel of $h^\star_{\text{CvM}}$ in the sense of (C.2) and define $\widetilde{h}^\star_{\text{CvM},c,d}$ in the sense of (C.4) for $0 \leq c, d, \leq 3$. Further denote the variance of $\widetilde{h}^\star_{\text{CvM},c,d}$ by $\sigma^2_{c,d}$ as in (C.6). Then the variance of U_{CvM} can be written as (Lemma C.1.3)

$$\text{Var}\,(U_{\text{CvM}}) = \sum_{c=0}^{3} \sum_{d=0}^{3} \frac{\binom{3}{c}\binom{3}{d}\binom{m-3}{3-c}\binom{n-3}{3-d}}{\binom{m}{3}\binom{n}{3}} \sigma^2_{c,d}. \tag{C.30}$$

First we bound $\sigma_{1,0}^2$. After applying the law of total expectation repeatedly, we can obtain that

$$\begin{aligned}
&\widetilde{h}^{\star}_{\mathrm{CvM},1,0}(x_1) - \mathbb{E}[\widetilde{h}^{\star}_{\mathrm{CvM},1,0}(x_1)] \\
=\ & \mathbb{E}\Big[\Big\{\mathbb{1}(\beta^\top x_1 \le \beta^\top X) - F_{\beta^\top X}(\beta^\top X)\Big\} \cdot \Big\{F_{\beta^\top Y}(\beta^\top X) - F_{\beta^\top X}(\beta^\top X)\Big\}\Big] \\
+\ & \mathbb{E}\Big[\Big\{\mathbb{1}(\beta^\top x_1 \le \beta^\top Y) - F_{\beta^\top X}(\beta^\top Y)\Big\} \cdot \Big\{F_{\beta^\top Y}(\beta^\top Y) - F_{\beta^\top X}(\beta^\top Y)\Big\}\Big] \\
+\ & \frac{1}{2}\mathbb{E}\Big[\Big\{F_{\beta^\top X}(\beta^\top x_1) - F_{\beta^\top Y}(\beta^\top x_1)\Big\}^2\Big] - \frac{1}{2}\mathbb{E}\Big[\Big\{F_{\beta^\top X}(\beta^\top X) - F_{\beta^\top Y}(\beta^\top X)\Big\}^2\Big] \\
=\ & f_1(x_1) + f_2(x_1) + f_3(x_1) \quad \text{(say)}.
\end{aligned}$$

Using the basic inequality $\{f_1(x_1) + f_2(x_1) + f_3(x_1)\}^2 \le 3f_1^2(x_1) + 3f_2^2(x_1) + 3f_3^2(x_1)$, we have

$$\begin{aligned}
\sigma_{1,0}^2 &= \mathbb{E}\big[\big\{\widetilde{h}^{\star}_{\mathrm{CvM},1,0}(X) - \mathbb{E}[\widetilde{h}^{\star}_{\mathrm{CvM},1,0}(X)]\big\}^2\big] \\
&\le 3\mathbb{E}\left[f_1^2(X)\right] + 3\mathbb{E}\left[f_2^2(X)\right] + 3\mathbb{E}\left[f_3^2(X)\right].
\end{aligned}$$

By applying Cauchy-Schwarz inequality, the first two terms are bounded by

$$\begin{aligned}
\mathbb{E}\left[f_1^2(X)\right] &\le \mathbb{E}\big[\big\{F_{\beta^\top X}(\beta^\top X) - F_{\beta^\top Y}(\beta^\top X)\big\}^2\big] \quad \text{and} \\
\mathbb{E}\left[f_2^2(X)\right] &\le \mathbb{E}\big[\big\{F_{\beta^\top X}(\beta^\top Y) - F_{\beta^\top Y}(\beta^\top Y)\big\}^2\big].
\end{aligned}$$

Since $0 \le \mathbb{E}\big[\big\{F_{\beta^\top X}(\beta^\top x_1) - F_{\beta^\top Y}(\beta^\top x_1)\big\}^2\big] \le 1$ for all $x_1 \in \mathbb{R}^d$, the third term is also bounded by

$$\begin{aligned}
\mathbb{E}\left[f_3^2(X)\right] &\le \frac{1}{4}\mathbb{E}\Big[\Big\{\mathbb{E}\big[\big\{F_{\beta^\top X}(\beta^\top X) - F_{\beta^\top Y}(\beta^\top X)\big\}^2\big]\Big\}^2\Big] \\
&\le \frac{1}{4}\mathbb{E}\big[\big\{F_{\beta^\top X}(\beta^\top X) - F_{\beta^\top Y}(\beta^\top X)\big\}^2\big].
\end{aligned}$$

Thus the following fact (see Theorem 4.1 of the main text)

$$\begin{aligned}
\mathbb{E}[U_{\mathrm{CvM}}] = &\frac{1}{2}\mathbb{E}\big[\big\{F_{\beta^\top X}(\beta^\top X) - F_{\beta^\top Y}(\beta^\top X)\big\}^2\big] \\
&+ \frac{1}{2}\mathbb{E}\big[\big\{F_{\beta^\top X}(\beta^\top Y) - F_{\beta^\top Y}(\beta^\top Y)\big\}^2\big],
\end{aligned}$$

leads to $\sigma_{1,0}^2 \lesssim \mathbb{E}[U_{\mathrm{CvM}}]$. Similarly we have $\sigma_{0,1}^2 \lesssim \mathbb{E}[U_{\mathrm{CvM}}]$. The rest of $\sigma_{c,d}^2$ can be uniformly bounded due to the boundedness of $\widetilde{h}^{\star}_{\mathrm{CvM}}$. Hence the result follows. □

Lemma C.1.12 (Two moments under permutations)**.** *The first and second moments of* U_{CvM} *under permutations are*

$$\mathbb{E}_{\varpi}\left[U_{\mathrm{CvM}}\right]=0 \quad \textit{and} \quad \mathbb{E}_{\varpi}\left[U_{\mathrm{CvM}}^{2}\right] \leq C\left(\frac{1}{m}+\frac{1}{n}\right)^{2},$$

where C *is a universal constant.*

Proof. Working directly with the kernel h_{CvM} is less intuitive to understand the moments of U_{CvM} under permutations. So we consider the third-order kernel $h_{\mathrm{CvM}}^{\star}$ in (C.9). Then from Lemma C.1.2, we have

$$U_{\mathrm{CvM}}=\frac{1}{(m)_3(n)_3}\sum_{i_1,i_2,i_3=1}^{m,\neq}\sum_{j_1,j_2,j_3=1}^{n,\neq} h_{\mathrm{CvM}}^{\star}(X_{i_1},X_{i_2},X_{i_3};Y_{j_1},Y_{j_2},Y_{j_3}).$$

1. First moment

Let $\{Z_1,\ldots,Z_{m+n}\}=\{X_1,\ldots,X_m,Y_1,\ldots,Y_n\}$ be the pooled samples. Then the first moment of U_{CvM} becomes

$$\mathbb{E}_{\varpi}\left[U_{\mathrm{CvM}}\right]=\mathbb{E}_{\varpi}\left[h_{\mathrm{CvM}}^{\star}(Z_{\varpi(1)},Z_{\varpi(2)},Z_{\varpi(3)};Z_{\varpi(m+1)},Z_{\varpi(m+2)},Z_{\varpi(m+3)})\right].$$

Notice that $h_{\mathrm{CvM}}^{\star}(x_1,x_2,x_3;y_1,y_2,y_3)=-h_{\mathrm{CvM}}^{\star}(y_1,x_2,x_3;x_1,y_2,y_3)$. This observation shows that the conditional expectation of $h_{\mathrm{CvM}}^{\star}$ given a subset of permutations $\mathcal{P}_{\varpi,4}=\{\varpi(2),\varpi(3),\varpi(m+2),\varpi(m+3)\}$ becomes zero, i.e.

$$\begin{aligned}\mathbb{E}_{\varpi(1),\varpi(m+1)}\Big[h_{\mathrm{CvM}}^{\star}\big(&Z_{\varpi(1)},Z_{\varpi(2)},Z_{\varpi(3)};\\ &Z_{\varpi(m+1)},Z_{\varpi(m+2)},Z_{\varpi(m+3)}\big)\big|\mathcal{P}_{\varpi,4}\Big]=0,\end{aligned}$$

for all $\mathcal{P}_{\varpi,4}$. Hence, $\mathbb{E}_{\varpi}\left[U_{\mathrm{CvM}}\right]=0$ by the law of total expectation.

2. Second moment

Next we calculate the second moment of U_{CvM} under permutations where

$$\begin{aligned}U_{\mathrm{CvM}}^{2}=&\frac{1}{(m)_3^2(n)_3^2}\times\\ &\sum_{i_1,i_2,i_3=1}^{m,\neq}\sum_{j_1,j_2,j_3=1}^{n,\neq}\sum_{i_1',i_2',i_3'=1}^{m,\neq}\sum_{j_1',j_2',j_3'=1}^{n,\neq}\Big\{h_{\mathrm{CvM}}^{\star}(Z_{i_1},Z_{i_2},Z_{i_3};Z_{j_1+m},Z_{j_2+m},Z_{j_3+m})\\ &\qquad\times h_{\mathrm{CvM}}^{\star}(Z_{i_1'},Z_{i_2'},Z_{i_3'};Z_{j_1'+m},Z_{j_2'+m},Z_{j_3'+m})\Big\}.\end{aligned}$$

Recall the definition of $\mathcal{I}_3$ and $\mathcal{J}_3$ given in (C.29). When $\#|\mathcal{I}_3| + \#|\mathcal{J}_3| \leq 1$, we apply the law of total expectation as in the proof of Lemma (C.1.10) to show that

$$\begin{aligned}\mathbb{E}_{\varpi}\big[&h^{\star}_{\mathrm{CvM}}(Z_{\varpi(i_1)}, Z_{\varpi(i_2)}, Z_{\varpi(i_3)}; Z_{\varpi(j_1+m)}, Z_{\varpi(j_2+m)}, Z_{\varpi(j_3+m)}) \\ &\times h^{\star}_{\mathrm{CvM}}(Z_{\varpi(i'_1)}, Z_{\varpi(i'_2)}, Z_{\varpi(i'_3)}; Z_{\varpi(j'_1+m)}, Z_{\varpi(j'_2+m)}, Z_{\varpi(j'_3+m)})\big] = 0.\end{aligned} \tag{C.31}$$

If $\#|\mathcal{I}_3| + \#|\mathcal{J}_3| > 1$, we use the fact that the kernel $h^{\star}_{\mathrm{CvM}}$ is bounded by one in absolute value to have

$$\begin{aligned}\big|\mathbb{E}_{\varpi}\big[&h^{\star}_{\mathrm{CvM}}(Z_{\varpi(i_1)}, Z_{\varpi(i_2)}, Z_{\varpi(i_3)}; Z_{\varpi(j_1+m)}, Z_{\varpi(j_2+m)}, Z_{\varpi(j_3+m)}) \\ &\times h^{\star}_{\mathrm{CvM}}(Z_{\varpi(i'_1)}, Z_{\varpi(i'_2)}, Z_{\varpi(i'_3)}; Z_{\varpi(j'_1+m)}, Z_{\varpi(j'_2+m)}, Z_{\varpi(j'_3+m)})\big]\big| \leq 1.\end{aligned}$$

Based on the previous observations and the fact that the size of the cases where $\#|\mathcal{I}_3| + \#|\mathcal{J}_3| > 1$ is at most $\prod_{i=0}^{4}(m-i) \times \prod_{j=0}^{6}(n-j) + \prod_{i=0}^{5}(m-i) \times \prod_{j=0}^{5}(n-j) + \prod_{i=0}^{6}(m-i) \times \prod_{j=0}^{4}(n-j)$ up to scaling factors, we conclude that

$$\mathbb{E}_{\varpi}\left[U^2_{\mathrm{CvM}}\right] \leq C\left(\frac{1}{m} + \frac{1}{n}\right)^2$$

as desired. □

Having established Lemma C.1.11 and Lemma C.1.12, we are now ready to prove that the CvM test is consistent under the contamination model.

1. Multivariate CvM-statistic

Note that since we assume that $Q_X \neq Q_Y$, there exists a positive constant δ_1 such that $W_d(P_{X,N}, P_{Y,N}) \geq (1-\epsilon)W_d(Q_X, Q_Y) \geq \delta_1$. Thus $\mathbb{E}[U_{\mathrm{CvM}}] \geq \delta_1^2$. We first upper bound the type II error as

$$\begin{aligned}\mathbb{P}_1\left(U_{\mathrm{CvM}} \leq c_{\alpha,\mathrm{CvM}}\right) =\ & \mathbb{P}_1\left(U_{\mathrm{CvM}} \leq c_{\alpha,\mathrm{CvM}}, c_{\alpha,\mathrm{CvM}} > \delta_1^2/2\right) \\ &+ \mathbb{P}_1\left(U_{\mathrm{CvM}} \leq c_{\alpha,\mathrm{CvM}}, c_{\alpha,\mathrm{CvM}} \leq \delta_1^2/2\right) \\ \leq\ & \mathbb{P}_1\left(c_{\alpha,\mathrm{CvM}} > \delta_1^2/2\right) + \mathbb{P}_1\left(U_{\mathrm{CvM}} \leq \delta_1^2/2\right) \\ =\ & (I) + (II) \quad \text{(say)}.\end{aligned}$$

For (I), Lemma C.1.12 and Chebyshev's inequality yield

$$\mathbb{P}_{\varpi}\left(U_{\mathrm{CvM}} \geq t\right) \leq \frac{\mathrm{Var}_{\varpi}(U_{\mathrm{CvM}})}{t^2} \leq \frac{C_0}{t^2} \cdot \left(\frac{1}{m} + \frac{1}{n}\right)^2$$

where C_0 is some universal constant. This shows that the critical value of the permutation test is uniformly bounded by

$$c_{\alpha,\mathrm{CvM}} \leq \sqrt{\frac{C_0}{\alpha}}\left(\frac{1}{m}+\frac{1}{n}\right).$$

Hence, we can bound (I) by

$$(I) = \mathbb{P}_1\left(c_{\alpha,\mathrm{CvM}} > \delta_1^2/2\right) \leq \frac{4}{\delta_1^4}\mathbb{E}_1\left[c_{\alpha,\mathrm{CvM}}^2\right] \leq \frac{4C_0}{\alpha\delta_1^4}\left(\frac{1}{m}+\frac{1}{n}\right)^2.$$

Next,

$$\begin{aligned}
(II) = \mathbb{P}_1\left(U_{\mathrm{CvM}} \leq \delta_1^2/2\right) &= \mathbb{P}_1\left(\frac{U_{\mathrm{CvM}}-\mathbb{E}_1[U_{\mathrm{CvM}}]}{\sqrt{\mathrm{Var}_1(U_{\mathrm{CvM}})}} \leq \frac{\delta_1^2/2-\mathbb{E}_1[U_{\mathrm{CvM}}]}{\sqrt{\mathrm{Var}_1(U_{\mathrm{CvM}})}}\right) \\
&\overset{(i)}{\leq} \mathbb{P}_1\left(\frac{U_{\mathrm{CvM}}-\mathbb{E}_1[U_{\mathrm{CvM}}]}{\sqrt{\mathrm{Var}_1(U_{\mathrm{CvM}})}} \leq \frac{-\delta_1^2/2}{\sqrt{\mathrm{Var}_1(U_{\mathrm{CvM}})}}\right) \\
&= \mathbb{P}_1\left(\frac{-U_{\mathrm{CvM}}+\mathbb{E}_1[U_{\mathrm{CvM}}]}{\sqrt{\mathrm{Var}_1(U_{\mathrm{CvM}})}} \geq \frac{\delta_1^2/2}{\sqrt{\mathrm{Var}_1(U_{\mathrm{CvM}})}}\right) \\
&\overset{(ii)}{\leq} \frac{4\mathrm{Var}_1(U_{\mathrm{CvM}})}{\delta_1^4} \\
&\overset{(iii)}{\leq} \frac{C_1}{\delta_1^2}\left(\frac{1}{m}+\frac{1}{n}\right)+\frac{C_2}{\delta_1^4}\left(\frac{1}{m}+\frac{1}{n}\right)^2
\end{aligned}$$

where (i) uses $\mathbb{E}[U_{\mathrm{CvM}}] \geq \delta_1^2$, (ii) is by Chebyshev's inequality and (iii) uses Lemma C.1.11 with universal constants C_1 and C_2. In the end, we have

$$\begin{aligned}
\lim_{m,n\to\infty}\inf_{G_N}\mathbb{E}_1[\phi_{\mathrm{CvM}}] \geq\ 1 - \lim_{m,n\to\infty}\inf_{G_N}\Bigg\{&\frac{4C_0}{\alpha\delta_1^4}\left(\frac{1}{m}+\frac{1}{n}\right)^2+\frac{C_1}{\delta_1^2}\left(\frac{1}{m}+\frac{1}{n}\right) \\
&+\frac{C_2}{\delta_1^4}\left(\frac{1}{m}+\frac{1}{n}\right)^2\Bigg\} = 1,
\end{aligned}$$

which completes the proof of the first part.

2. Energy statistic

Assume that G_N is a multivariate normal distribution with zero mean vector and covariance matrix $\sigma_N^2 I_d$ where $\sigma_N^2 \in \mathbb{R}$ is a positive sequence that tends to infinity as $N \to \infty$. Let us define the truncated random

vectors $\widetilde{X}$ and $\widetilde{Y}$ coupled with X and Y as

$$\widetilde{X} = \begin{cases} (0,\dots,0)^\top, & \text{if } X \sim Q_X, \\ X/\sigma_N, & \text{if } X \sim G_N, \end{cases} \quad \text{and} \widetilde{Y} = \begin{cases} (0,\dots,0)^\top, & \text{if } Y \sim Q_Y, \\ Y/\sigma_N, & \text{if } Y \sim G_N. \end{cases}$$

By construction, it is clear that $\widetilde{X}$ and $\widetilde{Y}$ have the same mixture distribution as

$$\widetilde{X}, \widetilde{Y} \sim \widetilde{P} := (1-\epsilon)Q_{\delta_0} + \epsilon \widetilde{G},$$

where Q_{δ_0} is the degenerate distribution at $(0,\dots,0)^\top$ and $\widetilde{G}$ is the standard multivariate normal distribution, i.e. $N((0,\dots,0)^\top, I_d)$. Now we consider the two energy statistics: one based on the original samples and the other based on the corresponding truncated samples. Denote these two statistics by U_{Energy} and $\widetilde{U}_{\text{Energy}}$, respectively. We shall first show that the energy statistic based on the original samples and the other energy statistic based on the truncated samples are asymptotically equivalent.

Lemma C.1.13. *Suppose $\sigma_N^2 \asymp N^q$ for some $q > 2$. Let $\widetilde{U}_{\text{Energy}}$ be the energy statistic based on $\{\widetilde{X}_1,\dots,\widetilde{X}_m,\widetilde{Y}_1,\dots,\widetilde{Y}_n\}$ coupled with the original samples $\{X_1,\dots,X_m,Y_1,\dots,Y_n\}$ and U_{Energy} be the energy statistic based on the original samples. Then under the asymptotic regime in (4.5),*

$$N\sigma_N^{-1}U_{\text{Energy}} - N\widetilde{U}_{\text{Energy}} \xrightarrow{p} 0.$$

Proof. Let us denote

$$\Delta_{m,n}(X_1,X_2) = \sigma_N^{-1}\|X_1 - X_2\| - \|\widetilde{X}_1 - \widetilde{X}_2\|.$$

Observe that there are four possible cases for $\Delta_{m,n}(X_1,X_2)$:

$$\Delta_{m,n}(X_1,X_2) = \begin{cases} \text{Case (a):} & \frac{1}{\sigma_N}\|X_1 - X_2\|, & \text{if } X_1, X_2 \sim Q_X, \\ \text{Case (b):} & \frac{1}{\sigma_N}\|X_1 - X_2\| - \frac{1}{\sigma_N}\|X_2\|, & \text{if } X_1 \sim Q_X, X_2 \sim G_N, \\ \text{Case (c):} & \frac{1}{\sigma_N}\|X_1 - X_2\| - \frac{1}{\sigma_N}\|X_1\|, & \text{if } X_1 \sim G_N, X_2 \sim Q_X, \\ \text{Case (d):} & 0, & \text{if } X_1, X_2 \sim H_m. \end{cases}$$

In any case, one can verify under the finite second moment condition that

$$\mathbb{E}\left[\Delta_{m,n}^2(X_1,X_2)\right] \lesssim \sigma_N^{-2}. \tag{C.32}$$

Similarly, it can be seen that $\mathbb{E}\left[\Delta_{m,n}^2(X_1,X_2)\right] \lesssim \sigma_N^{-2}$, $\mathbb{E}\left[\Delta_{m,n}^2(Y_1,Y_2)\right] \lesssim \sigma_N^{-2}$ and $\mathbb{E}\left[\Delta_{m,n}^2(X_1,Y_1)\right] \lesssim \sigma_N^{-2}$.

Let us write the symmetrized kernel of the energy statistic as

$$\begin{aligned}\widetilde{h}_{\text{Energy}}(x_1, x_2; y_1, y_2) &= \frac{1}{2}\|x_1 - y_1\| + \frac{1}{2}\|x_1 - y_1\| + \frac{1}{2}\|x_2 - y_1\| \\ &\quad + \frac{1}{2}\|x_2 - y_2\| - \|x_1 - x_2\| - \|y_1 - y_2\|.\end{aligned}$$

Then the energy statistic based on the truncated random samples can be written as

$$\widetilde{U}_{\text{Energy}} = \frac{1}{(m)_2(n)_2} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j_1,j_2=1}^{n,\neq} \widetilde{h}_{\text{Energy}}(\widetilde{X}_{i_1}, \widetilde{X}_{i_2}; \widetilde{Y}_{j_1}, \widetilde{Y}_{j_2}).$$

By letting

$$\begin{aligned}&h_D\{(X_{i_1}, \widetilde{X}_{i_1}), (X_{i_2}, \widetilde{X}_{i_2}); (Y_{j_1}, \widetilde{Y}_{j_1}), (Y_{j_2}, \widetilde{Y}_{j_2})\} \\ :=\ &\frac{1}{\sigma_N}\widetilde{h}_{\text{Energy}}(X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2}) - \widetilde{h}_{\text{Energy}}(\widetilde{X}_{i_1}, \widetilde{X}_{i_2}; \widetilde{Y}_{j_1}, \widetilde{Y}_{j_2}),\end{aligned} \tag{C.33}$$

the difference between the two energy statistics is

$$\begin{aligned}&N\big(\sigma_N^{-1}U_{\text{Energy}} - \widetilde{U}_{\text{Energy}}\big) \\ =\ &\frac{N}{(m)_2(n)_2} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j_1,j_2=1}^{n,\neq} h_D\{(X_{i_1}, \widetilde{X}_{i_1}), (X_{i_2}, \widetilde{X}_{i_2}); (Y_{j_1}, \widetilde{Y}_{j_1}), (Y_{j_2}, \widetilde{Y}_{j_2})\}.\end{aligned}$$

For simplicity we further write

$$h_D(i_1, i_2; j_1, j_2) = h_D\{(X_{i_1}, \widetilde{X}_{i_1}), (X_{i_2}, \widetilde{X}_{i_2}); (Y_{j_1}, \widetilde{Y}_{j_1}), (Y_{j_2}, \widetilde{Y}_{j_2})\}.$$

To show $N\big(\sigma_N^{-1}U_{\text{Energy}} - \widetilde{U}_{\text{Energy}}\big) \xrightarrow{p} 0$, we shall prove that the second moment of the difference converges to zero. To this end, we first apply Cauchy-Schwarz inequality to bound

$$\begin{aligned}&\mathbb{E}\big[h_D(i_1, i_2; j_1, j_2)h_D(i_1', i_2'; j_1', j_2')\big] \\ \le\ &\sqrt{\mathbb{E}\big[h_D^2(i_1, i_2; j_1, j_2)\big]}\sqrt{\mathbb{E}\big[h_D^2(i_1', i_2'; j_1', j_2')\big]}, \\ \lesssim\ &\sigma_N^{-2},\end{aligned}$$

which holds for any set of indices such that $i_1 \neq i_2, j_1 \neq j_2, i_1' \neq i_2', j_1' \neq j_2'$. Note that for the second inequality, we used

$$\mathbb{E}\big[h_D^2(i_1, i_2; j_1, j_2))\big] \lesssim\ \mathbb{E}[\Delta_{m,n}^2(X_{i_1}, X_{i_2})] + \mathbb{E}[\Delta_{m,n}^2(X_{i_1}, Y_{j_1})]$$

$$
\begin{aligned}
&+ \mathbb{E}[\Delta^2_{m,n}(X_{i_1}, Y_{j_2})] + \mathbb{E}[\Delta^2_{m,n}(X_{i_2}, Y_{i_1})] \\
&+ \mathbb{E}[\Delta^2_{m,n}(X_{i_2}, Y_{j_2})] + \mathbb{E}[\Delta^2_{m,n}(Y_{j_1}, Y_{j_2})], \\
&\lesssim \sigma_N^{-2},
\end{aligned}
$$

by the bound (C.32) and similarly for the other cases. As a consequence,

$$
\mathbb{E}\Big[N^2 \left(\sigma_N^{-1} U_{\text{Energy}} - \widetilde{U}_{\text{Energy}}\right)^2\Big] \lesssim \sigma_N^{-2} N^2.
$$

Under the given assumptions that $\sigma_N^2 \asymp (m+n)^q$ with $q > 2$ and $m/N \to \vartheta_X \in (0,1)$, we obtain $N(\sigma_N^{-1} U_{\text{Energy}} - \widetilde{U}_{\text{Energy}}) \xrightarrow{p} 0$ as desired. □

Since $\widetilde{U}_{\text{Energy}}$ has degeneracy of order one, $N\widetilde{U}_{\text{Energy}}$ converges to an infinite weighted sum of chi-square random variables (Theorem 4.2):

$$
N\widetilde{U}_{\text{Energy}} \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (\xi_k^2 - 1),
$$

for some $\{\lambda_k\}_{k=1}^{\infty}$. Lemma C.1.13 then implies that $NU_{\text{Energy}}/\sigma_N$ converges to the same distribution:

$$
\frac{N}{\sigma_N} U_{\text{Energy}} \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k (\xi_k^2 - 1).
$$

Furthermore, the permutation distribution of $N\sigma_N^{-1} U_{\text{Energy}}$ is asymptotically equivalent to the limiting distribution of $N\widetilde{U}_{\text{Energy}}$ as shown in the next lemma.

Lemma C.1.14. *Consider the same assumptions and notation used in Lemma C.1.13. Let $R(t)$ be the cumulative distribution function of the limiting distribution of $N\widetilde{U}_{\text{Energy}}$. Then the permutation distribution function of $N\sigma_N^{-1} U_{\text{Energy}}$, denoted by $\widehat{R}_{m,n}(t)$, satisfies*

$$
\sup_{t \in \mathbb{R}} \left| \widehat{R}_{m,n}(t) - R(t) \right| \xrightarrow{p} 0. \tag{C.34}
$$

Proof. Recall that $\{Z_1, \ldots, Z_{m+n}\}$ are the pooled samples of the original observations and we denote similarly by $\{\widetilde{Z}_1, \ldots, \widetilde{Z}_{m+n}\}$ the pooled samples of $\{\widetilde{X}_1, \ldots, \widetilde{X}_m, \widetilde{Y}_1, \ldots, \widetilde{Y}_n\}$. For any random permutation $\varpi = \{\varpi(1), \ldots, \varpi(N)\}$ of $\{1, \ldots, N\}$, we will show that

$$
N\sigma_N^{-1} U_{\text{Energy}}(Z_\varpi) - N\widetilde{U}_{\text{Energy}}(\widetilde{Z}_\varpi) \xrightarrow{p} 0, \tag{C.35}
$$

where $Z_\varpi = (Z_{\varpi(1)}, \ldots, Z_{\varpi(N)})$ and $\widetilde{Z}_\varpi = (\widetilde{Z}_{\varpi(1)}, \ldots, \widetilde{Z}_{\varpi(N)})$. If this is the case, then for two independent ϖ and ϖ', the following result

$$(N\widetilde{U}_{\text{Energy}}(\widetilde{Z}_\varpi), N\widetilde{U}_{\text{Energy}}(\widetilde{Z}_{\varpi'})) \xrightarrow{d} (U, U') \tag{C.36}$$

implies

$$(N\sigma_N^{-1}U_{\text{Energy}}(Z_\varpi), N\sigma_N^{-1}U_{\text{Energy}}(Z_{\varpi'})) \xrightarrow{d} (U, U'),$$

by Slutsky's theorem. Here U and U' are independent and identically distributed with the distribution function $R(t)$. Then Hoeffding's condition in Lemma C.1.4 establishes the convergence result (C.34). Indeed, (C.36) holds from Theorem C.1; hence it is enough to show (C.35) to complete the proof.

Note that

$$\begin{aligned}
& N\sigma_N^{-1}U_{\text{Energy}}(Z_\varpi) - N\widetilde{U}_{\text{Energy}}(\widetilde{Z}_\varpi) \\
= \; & \frac{N}{(m)_2(n)_2} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j_1,j_2=1}^{n,\neq} h_D\Big\{(Z_{\varpi(i_1)}, \widetilde{Z}_{\varpi(i_1)}), (Z_{\varpi(i_2)}, \widetilde{Z}_{\varpi(i_2)}); \\
& \qquad\qquad (Z_{\varpi(j_1+m)}, \widetilde{Z}_{\varpi(j_1+m)}), (Z_{\varpi(j_2+m)}, \widetilde{Z}_{\varpi(j_2+m)})\Big\},
\end{aligned}$$

where kernel h_D is given in (C.33). Note further by (C.32) that

$$\begin{aligned}
& \mathbb{E}\Big[h_D^2\Big\{(Z_{\varpi(i_1)}, \widetilde{Z}_{\varpi(i_1)}), (Z_{\varpi(i_2)}, \widetilde{Z}_{\varpi(i_2)}); \\
& \qquad (Z_{\varpi(j_1+m)}, \widetilde{Z}_{\varpi(j_1+m)}), (Z_{\varpi(j_2+m)}, \widetilde{Z}_{\varpi(j_2+m)})\Big\}\Big] \\
\lesssim \; & \mathbb{E}\left[\Delta_{m,n}^2(Z_{\varpi(i_1)}, Z_{\varpi(i_2)})\right] + \mathbb{E}\left[\Delta_{m,n}^2(Z_{\varpi(i_1)}, Z_{\varpi(j_1+m)})\right] \\
& + \mathbb{E}\left[\Delta_{m,n}^2(Z_{\varpi(i_1)}, Z_{\varpi(j_2+m)})\right] + \mathbb{E}\left[\Delta_{m,n}^2(Z_{\varpi(i_2)}, Z_{\varpi(j_1+m)})\right] \\
& + \mathbb{E}\left[\Delta_{m,n}^2(Z_{\varpi(i_2)}, Z_{\varpi(j_2+m)})\right] + \mathbb{E}\left[\Delta_{m,n}^2(Z_{\varpi(j_1+m)}, Z_{\varpi(j_2+m)})\right] \\
\lesssim \; & \sigma_N^{-2}
\end{aligned}$$

and similarly for the other cases. Then it is easy to see that

$$\mathbb{E}\big[\big(N\sigma_N^{-1}U_{\text{Energy}}(Z_\varpi) - N\widetilde{U}_{\text{Energy}}(\widetilde{Z}_\varpi)\big)^2\big] \lesssim \sigma_N^{-2}N^2 = o(1),$$

whenever $\sigma_N^2 \asymp N^q$ for some $q > 2$. This implies (C.35), which completes the proof. □

Combining the previous results yields

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}\left(U_{\text{Energy}} > c_{\alpha,\text{Energy}}\right) &= \lim_{N\to\infty} \mathbb{P}\left(N\sigma_N^{-1} U_{\text{Energy}} > N\sigma_N^{-1} c_{\alpha,\text{Energy}}\right) \\ &= \lim_{N\to\infty} \mathbb{P}\left(N\widetilde{U}_{\text{Energy}} > \widetilde{c}_{\alpha,\text{Energy}}\right) \leq \alpha,\end{aligned}$$

where $\widetilde{c}_{\alpha,\text{Energy}}$ is the $(1-\alpha)$ quantile of the permutation distribution of $N\widetilde{U}_{\text{Energy}}$. Hence the result follows.

C.4.12 Proof of Theorem 4.8

The minimax lower bound is based on a standard application of Neyman–Pearson lemma (see e.g. Baraud, 2002). Here we write the joint distributions of samples under the null and alternative hypotheses by $P_0^{m,n}$ and $P_1^{m,n}$, respectively. Then

$$\begin{aligned}\inf_{\phi\in\mathbb{T}_{m,n}(\alpha)} \sup_{P_X,P_Y\in\mathcal{F}(\epsilon^\star_{m,n})} \mathbb{P}_1\left(\phi = 0\right) &\geq 1-\alpha - \sup_{A\in\mathcal{A}} \left|P_0^{m,n}(A) - P_1^{m,n}(A)\right| \\ &\geq 1-\alpha - \sqrt{\frac{1}{2}\mathsf{KL}\left(P_1^{m,n}, P_0^{m,n}\right)}, \qquad \text{(C.37)}\end{aligned}$$

where the second inequality is by Pinsker's inequality (e.g. Lemma 2.5 of Tsybakov, 2009).

Recall the example considered in Lemma C.1.7:

$$X^* := (\xi_1, 0, \ldots, 0)^\top \quad \text{and} \quad Y^* := (\xi_2, 0, \ldots, 0)^\top,$$

where $\xi_1 \sim N(\mu_{X^*}, 1)$ and $\xi_2 \sim N(\mu_{Y^*}, 1)$. We let $\mu_{X^*} = \mu_{Y^*} = 0$ under the null and

$$\mu_{X^*} = \frac{\sqrt{2}(1-\alpha-\zeta)}{\sqrt{m}} \quad \text{and} \quad \mu_{Y^*} = -\frac{\sqrt{2}(1-\alpha-\zeta)}{\sqrt{n}},$$

under the alternative. Then from Lemma C.1.7, we have $P_{X^*}, P_{Y^*} \in \mathcal{F}(\epsilon^*_{m,n})$ for all d. In this case, the Kullback-Leibler divergence is calculated as

$$\mathsf{KL}\left(P_1^{m,n}, P_0^{m,n}\right) = \frac{m}{2}\mu_{X^*}^2 + \frac{n}{2}\mu_{Y^*}^2 = 2(1-\alpha-\zeta)^2.$$

By plugging this into (C.37), we conclude that

$$\inf_{\phi\in\mathbb{T}_{m,n}(\alpha)} \sup_{P_X,P_Y\in\mathcal{F}(\epsilon^\star_{m,n})} \mathbb{P}_1\left(\phi = 0\right) \geq \zeta.$$

Hence the result follows.

C.4.13 Proof of Theorem 5.5

Note that the permutation critical value $c_{\alpha,\mathrm{CvM}}$ is a random quantity depending on $\mathcal{X}_m$ and $\mathcal{Y}_n$. To control the randomness from $c_{\alpha,\mathrm{CvM}}$, we use a similar idea in Fromont et al. (2013) (see also Albert, 2015) where they considered the quantile of a permutation critical value. Specifically, let $c^*_{\zeta/2}$ be the upper $\zeta/2$ quantile of the distribution of $c_{\alpha,\mathrm{CvM}}$, and let Var_1 be the variance under H_1. Then it suffices to show that

$$\mathbb{E}_1\left[U_{\mathrm{CvM}}\right] \geq c^*_{\zeta/2} + \sqrt{\frac{2}{\zeta}\mathrm{Var}_1(U_{\mathrm{CvM}})} \tag{C.38}$$

uniformly over $P_X, P_Y \in \mathcal{F}(\epsilon^\star_{m,n})$ by choosing a sufficiently large c. In detail, we have

$$\begin{aligned}
&\mathbb{P}_1\left(U_{\mathrm{CvM}} < c_{\alpha,\mathrm{CvM}}\right) \\
=\ &\mathbb{P}_1\left(U_{\mathrm{CvM}} < c_{\alpha,\mathrm{CvM}},\ c_{\alpha,\mathrm{CvM}} > c^*_{\zeta/2}\right) + \mathbb{P}_1\left(U_{\mathrm{CvM}} < c_{\alpha,\mathrm{CvM}},\ c_{\alpha,\mathrm{CvM}} \leq c^*_{\zeta/2}\right) \\
\leq\ &\mathbb{P}_1\left(c_{\alpha,\mathrm{CvM}} > c^*_{\zeta/2}\right) + \mathbb{P}_1\left(U_{\mathrm{CvM}} \leq c^*_{\zeta/2}\right) \\
\leq\ &\frac{\zeta}{2} + \mathbb{P}_1\left(U_{\mathrm{CvM}} \leq c^*_{\zeta/2}\right),
\end{aligned}$$

where the second inequality is by the definition of $c^*_{\zeta/2}$. To control the second term, we apply Chebyshev's inequality

$$\begin{aligned}
\mathbb{P}_1\left(U_{\mathrm{CvM}} \leq c^*_{\zeta/2}\right) =\ &\mathbb{P}_1\left(\frac{U_{\mathrm{CvM}} - \mathbb{E}_1\left[U_{\mathrm{CvM}}\right]}{\sqrt{\mathrm{Var}_1\left(U_{\mathrm{CvM}}\right)}} \leq \frac{c^*_{\zeta/2} - \mathbb{E}_1\left[U_{\mathrm{CvM}}\right]}{\sqrt{\mathrm{Var}_1(U_{\mathrm{CvM}})}}\right) \\
=\ &\mathbb{P}_1\left(\frac{-U_{\mathrm{CvM}} + \mathbb{E}_1\left[U_{\mathrm{CvM}}\right]}{\sqrt{\mathrm{Var}_1\left(U_{\mathrm{CvM}}\right)}} \geq \frac{\mathbb{E}_1\left[U_{\mathrm{CvM}}\right] - c^*_{\zeta/2}}{\sqrt{\mathrm{Var}_1(U_{\mathrm{CvM}})}}\right) \\
\leq\ &\frac{\mathrm{Var}_1\left(U_{\mathrm{CvM}}\right)}{\left(\mathbb{E}_1\left[U_{\mathrm{CvM}}\right] - c^*_{\zeta/2}\right)^2} \\
\leq\ &\frac{\zeta}{2},
\end{aligned}$$

where the last inequality uses (C.38). To finish the proof, we only need to verify the condition in (C.38). Using Chebyshev's inequality and Lemma C.1.12,

$$\mathbb{P}_\varpi\left(U_{\mathrm{CvM}} \geq t\right) \leq \frac{\mathbb{E}_\varpi[U^2_{\mathrm{CvM}}]}{t^2} \leq \frac{C_0}{t^2}\left(\frac{1}{m} + \frac{1}{n}\right)^2.$$

As a result, the permutation critical value $c_{\alpha,\mathrm{CvM}}$ is upper bounded by $\sqrt{C_0/\alpha}(1/m+1/n)$ with probability one. This implies that its $\zeta/2$ upper quantile $c^*_{\zeta/2}$ is also bounded by

$$c^*_{\zeta/2} \;\le\; \sqrt{\frac{C_0}{\alpha}}\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)^2.$$

From Lemma C.1.11, we have

$$\begin{aligned}\sqrt{\frac{\zeta}{2}\mathrm{Var}_1\left[U_{\mathrm{CvM}}\right]} \;&\le\; \sqrt{\frac{\zeta}{2}\cdot\left\{C_1\mathbb{E}_1\left[U_{\mathrm{CvM}}\right]\cdot\left(\frac{1}{m}+\frac{1}{n}\right)+\frac{C_2}{m^2}+\frac{C_3}{n^2}+\frac{C_4}{mn}\right\}}\\ &\le\; C_5\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)^2.\end{aligned}$$

By choosing a sufficiently large $c>0$ in (4.15), we conclude that

$$\mathbb{E}_1[U_{\mathrm{CvM}}] \ge c^*_{\zeta/2}+\sqrt{\frac{\zeta}{2}\mathrm{Var}_1\left[U_{\mathrm{CvM}}\right]}.$$

This completes the proof of Theorem 5.5.

C.4.14 Proof of Proposition 4.1

Consider $P_{X,N}=(1-\epsilon)Q_X+\epsilon G_N, P_{Y,N}=(1-\epsilon)Q_Y+\epsilon G_N$ in (4.11) where Q_X and Q_Y are fixed but $Q_X\neq Q_Y$ and they have their finite second moments. Then as noted in Remark C.2, there exists a constant $\delta>0$ such that $W_d(P_{X,N},P_{Y,N})>\delta$. In other words, $P_{X,N},P_{Y,N}\in\mathcal{F}(\epsilon^\star_{m,n})$. Then the result follows by Theorem 4.7.

C.4.15 Proof of Theorem 4.10

The proof consists of two parts. In the first part, we will present some lemmas, which investigate the limiting behavior of $\widetilde{h}_{\mathrm{CvM}}$ under the HDLSS setting, and in part two, we will prove the main result.

- **Part 1.**

First define the five quantities

$$Q_1 := \frac{1}{3}-\frac{1}{2\pi}\arccos\left(\frac{\overline{\delta}^2_{XY}+\overline{\sigma}^2_X}{\overline{\delta}^2_{XY}+\overline{\sigma}^2_X+\overline{\sigma}^2_Y}\right)-\frac{1}{2\pi}\arccos\left(\frac{\overline{\delta}^2_{XY}+\overline{\sigma}^2_Y}{\overline{\delta}^2_{XY}+\overline{\sigma}^2_X+\overline{\sigma}^2_Y}\right),$$

$$
\begin{aligned}
Q_2 :=\ & \frac{1}{3} - \frac{1}{2\pi}\arccos\left(\frac{\overline{\sigma}_X^2}{(2\overline{\sigma}_X^2)^{1/2}(\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2)^{1/2}}\right) \\
& - \frac{1}{2\pi}\arccos\left(\frac{\overline{\sigma}_Y^2}{(2\overline{\sigma}_Y^2)^{1/2}(\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2)^{1/2}}\right),
\end{aligned}
$$

$$
\begin{aligned}
Q_3 :=\ & \frac{1}{3} - \frac{1}{4\pi}\left[\arccos\left(\frac{1}{2}\right) + \arccos\left(\frac{\overline{\delta}_{XY}^2 + \overline{\sigma}_Y^2}{\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2}\right)\right. \\
& \left. + 2\arccos\left(\frac{\overline{\sigma}_X^2}{(2\overline{\sigma}_X^2)^{1/2}(\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2)^{1/2}}\right)\right],
\end{aligned}
$$

$$
\begin{aligned}
Q_4 :=\ & \frac{1}{3} - \frac{1}{4\pi}\left[\arccos\left(\frac{1}{2}\right) + \arccos\left(\frac{\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2}{\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2}\right)\right. \\
& \left. + 2\arccos\left(\frac{\overline{\sigma}_Y^2}{(2\overline{\sigma}_Y^2)^{1/2}(\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2)^{1/2}}\right)\right],
\end{aligned}
$$

$$
Q_5 :=\ 0.
$$

Then by the weak law of large number and the continuous mapping theorem under **(A1)** and **(A2)**, it is not difficult to see that for any distinct indices $1 \le i_1, i_2, i_3, i_4 \le m$ and $1 \le j_1, j_2, j_3, i_4 \le n$,

$$
\begin{aligned}
\widetilde{h}_{\mathrm{CvM}}(X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2}) = \widetilde{h}_{\mathrm{CvM}}(Y_{j_1}, Y_{j_2}; X_{i_1}, X_{i_2}) \xrightarrow{p} Q_1, \\
\widetilde{h}_{\mathrm{CvM}}(X_{i_1}, Y_{j_1}; X_{i_2}, Y_{j_2}) = \widetilde{h}_{\mathrm{CvM}}(Y_{j_1}, X_{i_1}; Y_{j_2}, X_{i_2}) \xrightarrow{p} Q_2.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& \widetilde{h}_{\mathrm{CvM}}(X_{i_1}, X_{i_2}; X_{i_3}, Y_{j_1}) = \widetilde{h}_{\mathrm{CvM}}(X_{i_1}, X_{i_2}; Y_{j_1}, X_{i_3}) \\
= \ & \widetilde{h}_{\mathrm{CvM}}(X_{i_3}, Y_{j_1}; X_{i_1}, X_{i_2}) = \widetilde{h}_{\mathrm{CvM}}(Y_{j_1}, X_{i_3}; X_{i_1}, X_{i_2}) \xrightarrow{p} Q_3,
\end{aligned}
$$

and

$$
\begin{aligned}
& \widetilde{h}_{\mathrm{CvM}}(Y_{j_1}, Y_{j_2}; Y_{j_3}, X_{i_1}) = \widetilde{h}_{\mathrm{CvM}}(Y_{j_1}, Y_{j_2}; X_{i_1}, Y_{j_3}) \\
= \ & \widetilde{h}_{\mathrm{CvM}}(Y_{j_3}, X_{i_1}; Y_{j_1}, Y_{j_2}) = \widetilde{h}_{\mathrm{CvM}}(X_{i_1}, Y_{j_3}; Y_{j_1}, Y_{j_2}) \xrightarrow{p} Q_4.
\end{aligned}
$$

When all components are from the same distribution, then

$$\widetilde{h}_{\mathrm{CvM}}(X_{i_1}, X_{i_2}; X_{i_3}, X_{i_4}) \xrightarrow{p} Q_5 = 0 \quad \text{and}$$

$$\widetilde{h}_{\mathrm{CvM}}(Y_{j_1}, Y_{j_2}; Y_{j_3}, Y_{j_4}) \xrightarrow{p} Q_5 = 0.$$

In the next lemmas, we show that Q_1 is strictly greater than any of Q_2, Q_3, Q_4 and Q_5 whenever $\overline{\delta}^2_{XY} > 0$ or $\overline{\sigma}^2_X \neq \overline{\sigma}^2_Y$. In addition they all become equivalent to each other only when $\overline{\delta}^2_{XY} = 0$ and $\overline{\sigma}^2_X = \overline{\sigma}^2_Y$. We start by proving that the inverse cosine function is concave on $x \in [0, 1]$.

Lemma C.1.15. *The inverse cosine function is concave on $x \in [0, 1]$.*

Proof. The result follows by observing that

$$\frac{d}{dx}\arccos(x) = -\frac{1}{\sqrt{1-x^2}} \quad \text{and} \quad \frac{d^2}{dx^2}\arccos(x) = -\frac{x}{(1-x^2)^{3/2}}.$$

□

Lemma C.1.16. *Assume* **(A1)** *and* **(A2)** *hold. Then we have $Q_1 \geq Q_2$ and the equality holds if and only if $\overline{\delta}^2_{XY} = 0$ or $\overline{\sigma}^2_X = \overline{\sigma}^2_Y$.*

Proof. From Lemma C.1.15, the inverse cosine function is concave on $x \in [0, 1]$. So we apply reverse Jensen's inequality to have

$$\begin{aligned} &\arccos\left(\frac{\overline{\delta}^2_{XY} + \overline{\sigma}^2_X}{\overline{\delta}^2_{XY} + \overline{\sigma}^2_X + \overline{\sigma}^2_Y}\right) + \arccos\left(\frac{\overline{\delta}^2_{XY} + \overline{\sigma}^2_Y}{\overline{\delta}^2_{XY} + \overline{\sigma}^2_X + \overline{\sigma}^2_Y}\right) \\ \leq\ & 2\arccos\left(\frac{2\overline{\delta}^2_{XY} + \overline{\sigma}^2_X + \overline{\sigma}^2_Y}{2(\overline{\delta}^2_{XY} + \overline{\sigma}^2_X + \overline{\sigma}^2_Y)}\right). \end{aligned}$$

Then it is enough to show that

$$\begin{aligned} 2\arccos\left(\frac{2\overline{\delta}^2_{XY} + \overline{\sigma}^2_X + \overline{\sigma}^2_Y}{2(\overline{\delta}^2_{XY} + \overline{\sigma}^2_X + \overline{\sigma}^2_Y)}\right) \ \leq\ & \arccos\left(\frac{\overline{\sigma}^2_X}{(2\overline{\sigma}^2_X)^{1/2}(\overline{\delta}^2_{XY} + \overline{\sigma}^2_X + \overline{\sigma}^2_Y)^{1/2}}\right) \\ & + \arccos\left(\frac{\overline{\sigma}^2_Y}{(2\overline{\sigma}^2_Y)^{1/2}(\overline{\delta}^2_{XY} + \overline{\sigma}^2_X + \overline{\sigma}^2_Y)^{1/2}}\right). \end{aligned} \tag{C.39}$$

Before we proceed, we introduce the following quantities to simplify the expressions.

$$\begin{aligned} T_{XY} &= \frac{2\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2}{2(\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2)}, \\ T_X &= \frac{\overline{\sigma}_X^2}{(2\overline{\sigma}_X^2)^{1/2}(\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2)^{1/2}}, \\ T_Y &= \frac{\overline{\sigma}_Y^2}{(2\overline{\sigma}_Y^2)^{1/2}(\overline{\delta}_{XY}^2 + \overline{\sigma}_X^2 + \overline{\sigma}_Y^2)^{1/2}}, \end{aligned}$$

and

$$\begin{aligned} T_1 &= \overline{\delta}_{XY}^2(\overline{\sigma}_X^2 + 2\overline{\sigma}_Y^2 + 2\overline{\delta}_{XY}^2)^{1/2}\{2\overline{\sigma}_X^2 + \overline{\sigma}_Y^2 + 2\overline{\delta}_{XY}^2\}^{1/2}, \\ T_2 &= \overline{\delta}_{XY}^2(2\overline{\delta}_{XY}^2 - \overline{\sigma}_X\overline{\sigma}_Y), \\ T_3 &= (\overline{\sigma}_X^2 + \overline{\sigma}_Y^2)(\overline{\sigma}_X^2 + 2\overline{\sigma}_Y^2 + 2\overline{\delta}_{XY}^2)^{1/2}(2\overline{\sigma}_X^2 + \overline{\sigma}_Y^2 + 2\overline{\delta}_{XY}^2)^{1/2}, \\ T_4 &= -(\overline{\sigma}_X^2 + \overline{\sigma}_Y^2)(\overline{\sigma}_X^2 + \overline{\sigma}_Y^2 + \overline{\sigma}_X\overline{\sigma}_Y). \end{aligned}$$

Based on the monotonicity of the inverse cosine function and the basic identity

$$\arccos(x) + \arccos(y) = \arccos\big(xy - \sqrt{1-x^2}\sqrt{1-y^2}\big) \quad \text{for } 0 \le x, y \le 1,$$

it can be seen that proving the inequality (C.39) is equivalent to proving

$$2T_{XY}^2 - 1 \ge T_X T_Y - (1 - T_X^2)^{1/2}(1 - T_Y^2)^{1/2}. \tag{C.40}$$

After rearrangement, it can be further seen that the inequality (C.40) is equivalent to

$$T_1 + T_2 + T_3 + T_4 \ge 0. \tag{C.41}$$

The inequality (C.41) is indeed true and the equality holds only when $\overline{\delta}_{XY} = 0$ and $\overline{\sigma}_X^2 = \overline{\sigma}_Y^2$ since

$$\begin{aligned} &T_1 + T_2 \ge 0 \quad \textit{if and only if} \\ &\overline{\delta}_{XY}^4\{(6\overline{\sigma}_X^2 + 4\overline{\sigma}_X\overline{\sigma}_Y + 6\overline{\sigma}_Y^2)\overline{\delta}_{XY}^2 + 2(\overline{\sigma}_X^2 + \overline{\sigma}_Y^2)^2\} \ge 0, \end{aligned}$$

and

$$T_3 + T_4 \ge 0 \quad \textit{if and only if}$$

$$(\overline{\sigma}_X^2+\overline{\sigma}_Y^2)(\overline{\sigma}_X-\overline{\sigma}_Y)^2+2\overline{\delta}_{XY}^2(2\overline{\sigma}_X^2+\overline{\sigma}_Y^2)+2\overline{\delta}_{XY}^2(\overline{\sigma}_X^2+2\overline{\sigma}_Y^2)\geq 0.$$

This completes the proof. □

Lemma C.1.17. *Assume* **(A1)** *and* **(A2)** *hold. Then we have* $Q_1 \geq Q_3$ *and the equality holds if and only if* $\overline{\delta}_{XY}^2=0$ *or* $\overline{\sigma}_X^2=\overline{\sigma}_Y^2$.

Proof. Using reverse Jensen's inequality, we have

$$\arccos\left(\frac{1}{2}\right)\geq\frac{1}{2}\arccos\left(\frac{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2}{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}\right)+\frac{1}{2}\arccos\left(\frac{\overline{\delta}_{XY}^2+\overline{\sigma}_Y^2}{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}\right),$$

where the equality holds only when $\overline{\delta}_{XY}=0$ and $\overline{\sigma}_X^2=\overline{\sigma}_Y^2$. Then it is enough to verify that

$$\begin{aligned}&\arccos\left(\frac{\overline{\sigma}_X^2}{(2\overline{\sigma}_X^2)^{1/2}(\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2)^{1/2}}\right)\\ \geq\ &\frac{3}{4}\arccos\left(\frac{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2}{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}\right)+\frac{1}{4}\arccos\left(\frac{\overline{\delta}_{XY}^2+\overline{\sigma}_Y^2}{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}\right).\end{aligned}\tag{C.42}$$

By applying reverse Jensen's inequality and by the monotonicity of the inverse cosine function, it is seen that the following statement

$$\frac{4\overline{\delta}_{XY}^2+3\overline{\sigma}_X^2+\overline{\sigma}_Y^2}{4(\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2)}\ \geq\ \frac{\overline{\sigma}_X^2}{(2\overline{\sigma}_X^2)^{1/2}(\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2)^{1/2}}\tag{C.43}$$

implies (C.42). Since (C.43) is true if and only if

$$16\overline{\delta}_{XY}^4+16\overline{\delta}_{XY}^2\overline{\sigma}_X^2+8\overline{\delta}_{XY}^2\overline{\sigma}_Y^2+(\overline{\sigma}_X^2-\overline{\sigma}_Y^2)^2\geq 0\tag{C.44}$$

and the equality of (C.44) holds only if $\overline{\delta}_{XY}=0$ and $\overline{\sigma}_X^2=\overline{\sigma}_Y^2$, the result follows. □

Lemma C.1.18. *Assume* **(A1)** *and* **(A2)** *hold. Then we have* $Q_1 \geq Q_4$ *and the equality holds if and only if* $\overline{\delta}_{XY}^2=0$ *or* $\overline{\sigma}_X^2=\overline{\sigma}_Y^2$.

Proof. The proof is similar to that of Lemma C.1.17. Hence we omit the proof. □

Lemma C.1.19. *Assume* **(A1)** *and* **(A2)** *hold. Then we have* $Q_1 \geq Q_5$ *and the equality holds if and only if* $\overline{\delta}_{XY}^2=0$ *or* $\overline{\sigma}_X^2=\overline{\sigma}_Y^2$.

Proof. Using reverse Jensen's inequality, we see that

$$\begin{aligned}
&\frac{1}{\pi}\arccos\left(\frac{2\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}{2(\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2)}\right)\\
\geq\ &\frac{1}{2\pi}\arccos\left(\frac{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2}{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}\right)+\frac{1}{2\pi}\arccos\left(\frac{\overline{\delta}_{XY}^2+\overline{\sigma}_Y^2}{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}\right).
\end{aligned}$$

In addition, the inverse cosine function is monotone decreasing. So

$$\frac{1}{\pi}\arccos\left(\frac{2\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}{2(\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2)}\right)\leq\frac{1}{\pi}\arccos\left(\frac{\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2}{2(\overline{\delta}_{XY}^2+\overline{\sigma}_X^2+\overline{\sigma}_Y^2)}\right)=\frac{1}{3},$$

where the last step uses

$$\frac{1}{\pi}\arccos\left(\frac{1}{2}\right)=\frac{1}{3}.$$

Notice that the first inequality becomes the equality only when $\overline{\sigma}_X^2=\overline{\sigma}_Y^2$. The second inequality becomes the equality only when $\overline{\delta}_{XY}^2=0$. This proves the result. □

Combining the previous lemmas, we give a summary:

Lemma C.1.20. *Assume* **(A1)** *and* **(A2)** *hold. Then we have*

$$Q_1\geq\max\{Q_2,Q_3,Q_4,Q_5\}$$

and the equality holds as $Q_1=Q_2=Q_3=Q_4=Q_5$ *if and only if* $\overline{\delta}_{XY}^2=0$ *or* $\overline{\sigma}_X^2=\overline{\sigma}_Y^2$.

• **Part 2.**

In this part, we prove Theorem 4.10. Notice that U_{CvM} is a linear combination of kernel $\widetilde{h}_{\mathrm{CvM}}$ evaluated on different samples. Hence from the previous observation made in Part 1, it is seen that

$$U_{\mathrm{CvM}}\xrightarrow{p}Q_1\quad\text{under } H_1.$$

For a given permutation ϖ of $\{1,\ldots,N\}$, let us denote by $U_{\mathrm{CvM}}^{\varpi}$, the U-statistic computed based on $\{Z_{\varpi(1)},\ldots,Z_{\varpi(N)}\}$, i.e. $U_{\mathrm{CvM}}(Z_{\varpi(1)},\ldots,Z_{\varpi(N)})$. Let $\varpi_0=\{1,\ldots,N\}$ be the original permutation. Then $U_{\mathrm{CvM}}^{\varpi_0}$ becomes $U_{\mathrm{CvM}}(Z_1,\ldots,Z_N)$ computed based on the original samples. Let us define that the permutation ϖ is a neighbor of ϖ_0 if $\#|\{\varpi(1),\ldots,\varpi(m)\}\cap\{1,\ldots,m\}|=m$.

We first consider the unbalanced case where $m \neq n$. Observe that $U^{\varpi}_{\mathrm{CvM}}$ converges to Q_{ϖ}, which is a weighted average of $Q_1, \ldots, Q_5$. According to Lemma C.1.20, $Q_1 \geq Q_{\varpi}$ and it is not difficult to see that $Q_1 = Q_{\varpi}$ only if ϖ is a neighbor of ϖ_0. This means that $U^{\varpi_0}_{\mathrm{CvM}} > U^{\varpi}_{\mathrm{CvM}}$ in the limit for all ϖ but neighbors of ϖ_0 under H_1. Since there are $m!n!$ neighbors of ϖ_0 out of $N!$ permutations, if we choose $\alpha > 1/\{N!/(m!n!)\}$, then we have $\lim_{d\to\infty} \mathbb{E}[\phi_{\mathrm{CvM}}] = 1$.

For the balanced case where $m = n$, the result follows by a similar argument but now we also need to consider ϖ that satisfies $\#|\{\varpi(1), \ldots, \varpi(m)\} \cap \{m+1, \ldots, m+n\}| = n$ to be a neighbor of ϖ_0. This is because $U_{\mathrm{CvM}}(Z_1, \ldots, Z_N) = U_{\mathrm{CvM}}(Z_N, \ldots, Z_1)$ if $m = n$. Hence now we have $2m!n!$ neighbors of ϖ_0 out of $N!$ permutations and if we choose $\alpha > 2/\{N!/(m!n!)\}$, then we have $\lim_{d\to\infty} \mathbb{E}[\phi_{\mathrm{CvM}}] = 1$.

C.4.16 Proof of Theorem 4.11

Our strategy to prove the given result is to connect different statistics to the CQ statistic, which is relatively easy to handle. Each connection can be found in

- Section C.4.16: Connection of $U^{\varpi}_{\mathrm{CvM}}$ to U^{ϖ}_{CQ},
- Section C.4.16: Connection of $U^{\varpi}_{\mathrm{WMW}}$ to U^{ϖ}_{CQ},
- Section C.4.16: Connection of $U^{\varpi}_{\mathrm{Energy}}$ to U^{ϖ}_{CQ},
- Section C.4.16: Connection of $U^{\varpi}_{\mathrm{MMD}}$ to U^{ϖ}_{CQ}.

For notational simplicity, we will denote $Z_i^*, Z_2^*, Z_3^*, Z_4^*$ by Z_1, Z_2, Z_3, Z_4 throughout this section.

Connection of U^{ϖ}_{CvM} to U^{ϖ}_{CQ}

In this subsection, we connect $U^{\varpi}_{\mathrm{CvM}}$ to U^{ϖ}_{CQ} under the HDLSS setting. We first list some lemmas and their proofs. The final connection between $U^{\varpi}_{\mathrm{CvM}}$ and U^{ϖ}_{CQ} can be found in Proposition C.1.

Lemma C.1.21. *Under* **(A1)**, **(A2)** *and* **(A4)**, *we have*

$$\frac{1}{d}\|Z_1 - Z_2\|^2 - 2\overline{\sigma}_d^2 = O_{\mathbb{P}}(d^{-1/2}) \quad \textit{and}$$
$$\frac{1}{d}(Z_1 - Z_3)^{\top}(Z_2 - Z_3) = \overline{\sigma}_d^2 + O_{\mathbb{P}}(d^{-1/2}).$$

Proof. Under the assumption that $\mathrm{Var}[\|Z_1 - Z_2\|^2] = O(d)$, we apply Chebyshev's inequality to obtain

$$\frac{1}{d}\|Z_1 - Z_2\|^2 - \frac{1}{d}\mathbb{E}[\|Z_1 - Z_2\|^2] = O_{\mathbb{P}}(d^{-1/2}).$$

Note that regardless of the distributions of Z_1 and Z_2, the expected value of $\|Z_1 - Z_2\|^2$ is bounded by

$$\mathbb{E}[\|Z_1 - Z_2\|^2] \leq \|\mu_X - \mu_Y\|^2 + 2\mathrm{tr}(\Sigma^2).$$

Thus under **(A4)**,

$$\frac{1}{d}\mathbb{E}[\|Z_1 - Z_2\|^2] - 2\overline{\sigma}_d^2 = O(d^{-1/2}).$$

By combining the results, we prove the first part. The second part follows similarly. □

Lemma C.1.22. *Under* **(A1)**, **(A2)** *and* **(A4)**, *we have*

$$\frac{\sqrt{d}}{\|Z_1 - Z_2\|} = \frac{1}{(2\overline{\sigma}_d^2)^{1/2}} - \frac{1}{2(2\overline{\sigma}_d^2)^{3/2}}\left(d^{-1}\|Z_1 - Z_2\|^2 - 2\overline{\sigma}_d^2\right) + O_{\mathbb{P}}(d^{-1}).$$

Proof. Consider $f(x) = 1/\sqrt{x}$ and represent

$$f\left(d^{-1}\|Z_1 - Z_2\|^2\right) = \frac{\sqrt{d}}{\|Z_1 - Z_2\|}.$$

By using the second order Taylor expansion of $f(x)$ around $f(2\overline{\sigma}_d^2)$ with Lemma C.1.21, we obtain the result. □

Lemma C.1.23. *Under* **(A1)**, **(A2)** *and* **(A4)**, *we have*

$$\begin{aligned}\frac{d}{\|Z_1 - Z_3\|\|Z_2 - Z_3\|} = \frac{1}{2\overline{\sigma}_d^2} &- \frac{1}{8\overline{\sigma}_d^4}\left(d^{-1}\|Z_1 - Z_3\|^2 - 2\overline{\sigma}_d^2\right) \\ &- \frac{1}{8\overline{\sigma}_d^4}\left(d^{-1}\|Z_2 - Z_3\|^2 - 2\overline{\sigma}_d^2\right) + O_{\mathbb{P}}(d^{-1}).\end{aligned}$$

Proof. Based on Lemma C.1.22, we have

$$\begin{aligned}&\frac{d}{\|Z_1 - Z_3\|\|Z_2 - Z_3\|} \\ =\ &\left\{\frac{1}{(2\overline{\sigma}_d^2)^{1/2}} - \frac{1}{2(2\overline{\sigma}_d^2)^{3/2}}\left(d^{-1}\|Z_1 - Z_3\|^2 - 2\overline{\sigma}_d^2\right) + O_{\mathbb{P}}(d^{-1})\right\}\end{aligned}$$

$$\times \left\{ \frac{1}{(2\overline{\sigma}_d^2)^{1/2}} - \frac{1}{2(2\overline{\sigma}_d^2)^{3/2}} \left(d^{-1}\|Z_2 - Z_3\|^2 - 2\overline{\sigma}_d^2 \right) + O_{\mathbb{P}}(d^{-1}) \right\}.$$

By expanding the right-hand side and the following observations made from Lemma C.1.21,

$$\frac{1}{2(2\overline{\sigma}_d^2)^{3/2}} \left(d^{-1}\|Z_1 - Z_3\|^2 - 2\overline{\sigma}_d^2 \right) = O_{\mathbb{P}}(d^{-1/2}),$$
$$\frac{1}{2(2\overline{\sigma}_d^2)^{3/2}} \left(d^{-1}\|Z_2 - Z_3\|^2 - 2\overline{\sigma}_d^2 \right) = O_{\mathbb{P}}(d^{-1/2}),$$

the result follows. □

Lemma C.1.24. *Under* **(A1)**, **(A2)** *and* **(A4)**, *we have*

$$\begin{aligned} &\arccos\left\{ \frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{\|Z_1 - Z_3\|\|Z_2 - Z_3\|} \right\} \\ =\ &\arccos\left(\frac{1}{2}\right) - \frac{2}{\sqrt{3}} \left\{ \frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{\|Z_1 - Z_3\|\|Z_2 - Z_3\|} - \frac{1}{2} \right\} + O_{\mathbb{P}}(d^{-1}). \end{aligned}$$

Proof. First note that

$$\frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{\|Z_1 - Z_3\|\|Z_2 - Z_3\|} - \frac{1}{2} = O_{\mathbb{P}}(d^{-1/2}),$$

which follows from Lemma C.1.21 and Lemma C.1.23. We then use the second order Taylor expansion of the inverse cosine function around $\arccos(1/2)$ to obtain the result. □

Lemma C.1.25. *Under* **(A1)**, **(A2)** *and* **(A4)**, *we have*

$$\begin{aligned} &\frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{\|Z_1 - Z_3\|\|Z_2 - Z_3\|} - \frac{1}{2} \\ =\ &\frac{(Z_1 - Z_3)^\top (Z_2 - Z_3) - d\overline{\sigma}_d^2}{2d\overline{\sigma}_d^2} - \frac{1}{8d\overline{\sigma}_d^2} \left(\|Z_1 - Z_3\|^2 + \|Z_2 - Z_3\|^2 - 4d\overline{\sigma}_d^2 \right) \\ &+ O_{\mathbb{P}}(d^{-1}). \end{aligned}$$

Proof. We split the left-hand side into two terms:

$$\begin{aligned} \frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{\|Z_1 - Z_3\|\|Z_2 - Z_3\|} - \frac{1}{2} = &\frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{\|Z_1 - Z_3\|\|Z_2 - Z_3\|} - \frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{2d\overline{\sigma}_d^2} \\ &+ \frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{2d\overline{\sigma}_d^2} - \frac{1}{2}. \end{aligned}$$

Now it is enough to show that

$$\frac{(Z_1-Z_3)^\top(Z_2-Z_3)}{\|Z_1-Z_3\|\|Z_2-Z_3\|}-\frac{(Z_1-Z_3)^\top(Z_2-Z_3)}{2d\overline{\sigma}_d^2}$$
$$=-\frac{1}{8d\overline{\sigma}_d^2}\left(\|Z_1-Z_3\|^2+\|Z_2-Z_3\|^2-4d\overline{\sigma}_d^2\right)+O_{\mathbb{P}}(d^{-1}).$$

Note that

$$\begin{aligned}&\frac{(Z_1-Z_3)^\top(Z_2-Z_3)}{\|Z_1-Z_3\|\|Z_2-Z_3\|}-\frac{(Z_1-Z_3)^\top(Z_2-Z_3)}{2d\overline{\sigma}_d^2}\\ =\;&(Z_1-Z_3)^\top(Z_2-Z_3)\times\left(\frac{1}{\|Z_1-Z_3\|\|Z_2-Z_3\|}-\frac{1}{2d\overline{\sigma}_d^2}\right)\\ =\;&(I)\times(II)\quad\text{(say)}.\end{aligned}$$

From Lemma C.1.21 and Lemma C.1.23, it is seen that

$$(I)=d\overline{\sigma}_d^2+O_{\mathbb{P}}(d^{1/2}),$$
$$(II)=-\frac{1}{8d\overline{\sigma}_d^4}\left[d^{-1}\|Z_1-Z_3\|^2+d^{-1}\|Z_1-Z_3\|^2-4\overline{\sigma}_d^2+O_{\mathbb{P}}(d^{-2})\right].$$

Expanding the terms in $(I)\times(II)$, we obtain the result. □

Based on the previous lemmas, we prove the main result of this subsection.

Proposition C.1. *Under* **(A1)**, **(A2)** *and* **(A4)**, *we have*

$$\begin{aligned}&\widetilde{h}_{\mathrm{CvM}}(Z_1,Z_2;Z_3,Z_4)\\ =\;&\frac{1}{4\pi\sqrt{3}d\overline{\sigma}_d^2}\{(Z_1-Z_3)^\top(Z_2-Z_4)+(Z_1-Z_4)^\top(Z_2-Z_3)\}+O_{\mathbb{P}}(d^{-1})\end{aligned}\tag{C.45}$$

and thus

$$U_{\mathrm{CvM}}^{\varpi}=\frac{1}{2\pi\sqrt{3}d\overline{\sigma}_d^2}U_{\mathrm{CQ}}^{\varpi}+O_{\mathbb{P}}(d^{-1}).$$

Proof. By Lemma C.1.24 and Lemma C.1.25,

$$\arccos\left\{\frac{(Z_1-Z_3)^\top(Z_2-Z_3)}{\|Z_1-Z_3\|\|Z_2-Z_3\|}\right\}$$

$$= \arccos\left(\frac{1}{2}\right) - \frac{2}{\sqrt{3}}\left\{\frac{(Z_1 - Z_3)^\top (Z_2 - Z_3)}{2d\overline{\sigma}_d^2} - \frac{1}{2}\right.$$
$$\left. - \frac{1}{8d\overline{\sigma}_d^2}\left(\|Z_1 - Z_3\|^2 + \|Z_2 - Z_3\|^2 - 4d\overline{\sigma}_d^2\right)\right\} + O_{\mathbb{P}}(d^{-1}).$$

We can obtain (C.45) by first plugging the above approximation into $\widetilde{h}_{\mathrm{CvM}}$ for each inverse cosine function and then simplifying the expression. The second result is trivial by noting that

$$\widetilde{h}_{\mathrm{CQ}}(x_1, x_2; y_1, y_2) = \frac{1}{2}(x_1 - y_1)^\top (x_2 - y_2) + \frac{1}{2}(x_1 - y_2)^\top (x_2 - y_1)$$

is the symmetrized kernel of the CQ statistic. □

Connection of $U^{\varpi}_{\boldsymbol{WMW}}$ to $U^{\varpi}_{\boldsymbol{CQ}}$

Note that the symmetrized kernel of the WMW statistic can be written as

$$\widetilde{h}_{\mathrm{WMW}}(x_1, x_2; y_1, y_2) = \frac{1}{2}\frac{(x_1 - y_1)^\top (x_2 - y_2)}{\|x_1 - y_1\|\|x_2 - y_2\|} + \frac{1}{2}\frac{(x_1 - y_2)^\top (x_2 - y_1)}{\|x_1 - y_2\|\|x_2 - y_1\|}.$$

We first provide a couple of lemmas and their proofs. We then present the main result in Proposition C.2.

Lemma C.1.26. *Under* **(A1)**, **(A2)**, **(A3)** *and* **(A4)**, *we have*

$$\frac{d}{\|Z_1 - Z_2\|\|Z_3 - Z_4\|} = \frac{1}{2\overline{\sigma}_d^2} - \frac{1}{8\overline{\sigma}_d^4}\left(d^{-1}\|Z_1 - Z_2\|^2 - 2\overline{\sigma}_d^2\right)$$
$$- \frac{1}{8\overline{\sigma}_d^4}\left(d^{-1}\|Z_3 - Z_4\|^2 - 2\overline{\sigma}_d^2\right) + O_{\mathbb{P}}(d^{-1}).$$

The proof of this result is similar to Lemma C.1.23; hence omitted.

Lemma C.1.27. *Under* **(A1)**, **(A2)**, **(A3)** *and* **(A4)**, *we have*

$$\frac{(Z_1 - Z_3)^\top (Z_2 - Z_4)}{\|Z_1 - Z_3\|\|Z_2 - Z_4\|} = \frac{(Z_1 - Z_3)^\top (Z_2 - Z_4)}{2d\overline{\sigma}_d^2} + O_{\mathbb{P}}(d^{-1}).$$

Proof. Under **(A3)**, it can be seen as similar to Lemma C.1.21 that

$$d^{-1}(Z_1 - Z_3)^\top (Z_2 - Z_4) = O_{\mathbb{P}}(d^{-1/2}).$$

Then combining the above with Lemma C.1.21 and Lemma C.1.26,

$$\begin{aligned}
&\frac{(Z_1 - Z_3)^\top (Z_2 - Z_4)}{\|Z_1 - Z_3\|\|Z_2 - Z_4\|} - \frac{(Z_1 - Z_3)^\top (Z_2 - Z_4)}{2d\overline{\sigma}_d^2} \\
&= d^{-1}(Z_1 - Z_3)^\top (Z_2 - Z_4) \times \left\{ \frac{d}{\|Z_1 - Z_3\|\|Z_2 - Z_4\|} - \frac{1}{2\overline{\sigma}_d^2} \right\} \\
&= O_{\mathbb{P}}(d^{-1/2}) \times O_{\mathbb{P}}(d^{-1/2}).
\end{aligned}$$

Hence the result follows. □

Based on the previous lemmas, we prove the main result of this subsection.

Proposition C.2. *Under* **(A1)**, **(A2)**, **(A3)** *and* **(A4)**, *we have*

$$\begin{aligned}
&\widetilde{h}_{\mathrm{WMW}}(Z_1, Z_2; Z_3, Z_4) \\
&= \frac{1}{2d\overline{\sigma}_d^2}\{(Z_1 - Z_3)^\top (Z_2 - Z_4) + (Z_1 - Z_4)^\top (Z_2 - Z_3)\} + O_{\mathbb{P}}(d^{-1})
\end{aligned}$$

and thus

$$U_{\mathrm{WMW}} = \frac{1}{2d\overline{\sigma}_d^2} U_{\mathrm{CQ}} + O_{\mathbb{P}}(d^{-1}).$$

Proof. The result is a direct consequence of Lemma C.1.27. □

Connection of $U^{\varpi}_{\mathit{Energy}}$ to U^{ϖ}_{CQ}

Next we find a connection between $U^{\varpi}_{\mathrm{Energy}}$ and U^{ϖ}_{CQ}. Note that the symmetrized kernel of the energy statistic can be written as

$$\begin{aligned}
\widetilde{h}_{\mathrm{Energy}}(x_1, x_2; y_1, y_2) &= \frac{1}{2}\|x_1 - y_1\| + \frac{1}{2}\|x_1 - y_2\| + \frac{1}{2}\|x_2 - y_1\| \\
&\quad + \frac{1}{2}\|x_2 - y_2\| - \|x_1 - x_2\| - \|y_1 - y_2\|.
\end{aligned}$$

Using this kernel expression, we connect U_{Energy} to U_{CQ} in Proposition C.3.

We start with one lemma.

Lemma C.1.28. *Under* **(A1)** *and* **(A2)**, *we have*

$$\frac{1}{\sqrt{d}}\|Z_1 - Z_2\| = (2\overline{\sigma}_d^2)^{1/2} + \frac{1}{2(2\overline{\sigma}_d^2)^{1/2}}\left(d^{-1}\|Z_1 - Z_2\|^2 - 2\overline{\sigma}_d^2\right) + O_{\mathbb{P}}(d^{-1}).$$

Proof. We use the second order Taylor expansion of $f(x) = \sqrt{x}$ around $f(2\overline{\sigma}_d^2)$ with Lemma C.1.21 to prove this result. □

The main result of this subsection is stated as follows.

Proposition C.3. *Under* **(A1)** *and* **(A2)**, *we have*

$$\begin{aligned}
&\widetilde{h}_{\text{Energy}}(Z_1, Z_2; Z_3, Z_4) \\
=\ & \frac{1}{2(2d\overline{\sigma}_d^2)^{1/2}}\{(Z_1 - Z_3)^\top(Z_2 - Z_4) + (Z_1 - Z_4)^\top(Z_2 - Z_3)\} + O_{\mathbb{P}}(d^{-1/2})
\end{aligned}$$

and thus

$$U_{\text{Energy}} = \frac{1}{2(d\overline{\sigma}_d^2)^{1/2}}U_{\text{CQ}} + O_{\mathbb{P}}(d^{-1/2}).$$

Proof. We use Lemma C.1.28 to approximate $\widetilde{h}_{\text{Energy}}$ to $\widetilde{h}_{\text{CQ}}$ and simplify the expression to obtain the first result. The second result is trivial. □

Connection of U_{MMD}^{ϖ} to U_{CQ}^{ϖ}

In this subsection, we find a connection between U_{MMD}^{ϖ} and U_{CQ}^{ϖ}. The symmetrized kernel of the MMD statistic can be written as

$$\begin{aligned}
&\widetilde{h}_{\text{MMD}}(x_1, x_2; y_1, y_2) \\
=\ & -\frac{1}{2}\exp\left(-\frac{1}{2\varsigma_d^2}\|x_1 - y_1\|^2\right) - \frac{1}{2}\exp\left(-\frac{1}{2\varsigma_d^2}\|x_1 - y_2\|^2\right) \\
& -\frac{1}{2}\exp\left(-\frac{1}{2\varsigma_d^2}\|x_2 - y_1\|^2\right) - \frac{1}{2}\exp\left(-\frac{1}{2\varsigma_d^2}\|x_2 - y_2\|^2\right) \\
& +\exp\left(-\frac{1}{2\varsigma_d^2}\|x_1 - x_2\|^2\right) + \exp\left(-\frac{1}{2\varsigma_d^2}\|y_1 - y_2\|^2\right),
\end{aligned}$$

and we assume that $\varsigma_d^2 \asymp d$. We first provide an approximation of the Gaussian kernel.

Lemma C.1.29. *Under* **(A1)**, **(A2)** *and* $\varsigma_d^2 \asymp d$, *we have*

$$\begin{aligned}
&\exp\left(-\frac{1}{2\varsigma_d^2}\|Z_1 - Z_2\|^2\right) \\
=\ &\exp\left(-\frac{d\overline{\sigma}_d^2}{\varsigma_d^2}\right) - \exp\left(-\frac{d\overline{\sigma}_d^2}{\varsigma_d^2}\right)\left[\frac{1}{2\varsigma_d^2}\|Z_1 - Z_2\|^2 - \frac{d\overline{\sigma}_d^2}{\varsigma_d^2}\right] + O_{\mathbb{P}}(d^{-1}).
\end{aligned}$$

Proof. We consider the second order Taylor expansion of $f(x) = e^{-x}$ around $f(d\overline{\sigma}_d^2/\varsigma_d^2)$. Notice that under $\varsigma_d^2 \asymp d$, we have $d\overline{\sigma}_d^2/\varsigma_d^2 = O(1)$ and

$$\frac{1}{2\varsigma_d^2}\|Z_1 - Z_2\|^2 - \frac{d\overline{\sigma}_d^2}{\varsigma_d^2} = \frac{d}{2\varsigma_d^2}\big(d^{-1}\|Z_1 - Z_2\|^2 - 2\overline{\sigma}_d^2\big) = O_{\mathbb{P}}(d^{-1/2})$$

from Lemma C.1.21. Thus the result follows. □

The main result of this subsection is stated as follows.

Proposition C.4. *Under* **(A1)**, **(A2)** *and* $\varsigma_d^2 \asymp d$, *we have*

$$\begin{aligned}
&\widetilde{h}_{\mathrm{MMD}}(Z_1, Z_2; Z_3, Z_4) \\
=\ &\frac{e^{-d\overline{\sigma}_d^2/\varsigma_d^2}}{2\varsigma_d^2}\{(Z_1 - Z_3)^\top(Z_2 - Z_4) + (Z_1 - Z_4)^\top(Z_2 - Z_3)\} + O_{\mathbb{P}}(d^{-1}),
\end{aligned}$$

and thus

$$U_{\mathrm{MMD}} = \varsigma_d^{-2} e^{-d\overline{\sigma}_d^2/\varsigma_d^2} U_{\mathrm{CQ}} + O_{\mathbb{P}}(d^{-1/2}).$$

Proof. We use Lemma C.1.29 to approximate $\widetilde{h}_{\mathrm{MMD}}$ to $\widetilde{h}_{\mathrm{CQ}}$ and simplify the expression to obtain the first result. The second result is trivial. □

• Main proof of Theorem 4.11.

By collecting the results in Proposition C.1, Proposition C.2, Proposition C.3 and Proposition C.4, it is easily checked that Theorem 4.11 holds and thus we complete the proof.

C.4.17 Proof of Corollary 4.11.1

We will only show that $c_{\alpha,\mathrm{CvM}} = c_{\alpha,\mathrm{CQ}} + O_{\mathbb{P}}(d^{-1/2})$. The remaining results follow similarly. From Theorem 4.11, we know that

$$2\pi\sqrt{3d}\overline{\sigma}_d^2(U_{\mathrm{CvM}}^{\varpi_1},\ldots,U_{\mathrm{CvM}}^{\varpi_{N!}}) = d^{-1/2}(U_{\mathrm{CQ}}^{\varpi_1},\ldots,U_{\mathrm{CQ}}^{\varpi_{N!}}) + O_{\mathbb{P}}(d^{-1/2})$$

where ϖ_i is an element of $\mathcal{S}_N$ for $i = 1,\ldots,N!$. For simplicity, let us write $2\pi\sqrt{3d}\overline{\sigma}_d^2 U_{\mathrm{CvM}}^{\varpi_i} = U_{\mathrm{CvM},s}^{\varpi_i}$ and $d^{-1/2}U_{\mathrm{CQ}}^{\varpi_i} = U_{\mathrm{CQ},s}^{\varpi_i}$. Then $c_{\alpha,\mathrm{CvM}}$ and $c_{\alpha,\mathrm{CQ}}$ are the $\lceil N!(1-\alpha)\rceil$th order statistic of $\{U_{\mathrm{CvM},s}^{\varpi_1},\ldots,U_{\mathrm{CvM},s}^{\varpi_{N!}}\}$ and $\{U_{\mathrm{CQ},s}^{\varpi_1},\ldots,U_{\mathrm{CQ},s}^{\varpi_{N!}}\}$, respectively. It is well-known that the order statistic is a Lipschitz function (e.g. page 43 of Wainwright, 2019). More specifically, using Pigeonhole principle, it can be seen that

$$|c_{\alpha,\mathrm{CvM}} - c_{\alpha,\mathrm{CQ}}| \leq \left\{\sum_{i=1}^{N!}(U_{\mathrm{CvM},s}^{\varpi_i} - U_{\mathrm{CQ},s}^{\varpi_i})^2\right\}^{1/2} = O_{\mathbb{P}}(d^{-1/2}).$$

Hence the result follows.

C.4.18 Proof of Proposition 4.2

From the definition of ρ_{Angle}, it is seen that

$$\begin{aligned}
&2\mathbb{E}\left[\rho_{\mathrm{Angle}}(X_1,Y_1)\right] - \mathbb{E}\left[\rho_{\mathrm{Angle}}(X_1,X_2)\right] - \mathbb{E}\left[\rho_{\mathrm{Angle}}(Y_1,Y_2)\right] \\
=\ &\frac{1}{\pi}\mathbb{E}\left[\mathsf{Ang}(X_1-X_2,Y_1-X_2)\right] + \frac{1}{\pi}\mathbb{E}\left[\mathsf{Ang}(X_1-Y_2,Y_1-Y_2)\right] \\
&-\frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}(X_1-X_3,X_2-X_3)\right] - \frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}(X_1-Y_1,X_2-Y_1)\right] \\
&-\frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}(Y_1-X_2,Y_2-X_2)\right] - \frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}(Y_1-Y_3,Y_2-Y_3)\right].
\end{aligned}$$

Then the result follows by Lemma C.1.1.

C.4.19 Proof of Proposition 4.3

Given $\beta \in \mathbb{S}^{d-1}$, note that

$$\left(\mathbb{P}(\beta^\top Z_1 > 0) - \frac{1}{2}\right)^2 = \frac{1}{4} - \mathbb{E}\left[\mathbb{1}(\beta^\top Z_1 > 0)\right] + \mathbb{E}\left[\mathbb{1}(\beta^\top Z_1 > 0)\mathbb{1}(\beta^\top Z_2 > 0)\right].$$

Applying Lemma 4.0.2 with Fubini's theorem yields

$$\mathbb{E}\left[\int_{\mathbb{S}^{d-1}} \mathbb{1}(\beta^\top Z_1 > 0) d\lambda(\beta)\right] = \frac{1}{2},$$
$$\mathbb{E}\left[\int_{\mathbb{S}^{d-1}} \mathbb{1}(\beta^\top Z_1 > 0)\mathbb{1}(\beta^\top Z_2 > 0) d\lambda(\beta)\right] = \frac{1}{2} - \frac{1}{2\pi}\mathbb{E}\left[\mathsf{Ang}\left(Z_1, Z_2\right)\right].$$

This completes the proof.

C.4.20 Proof of Proposition 4.4

The result follows by replacing Z_1, Z_2 with $X_1 - Y_1$, $X_2 - Y_2$ in Proposition 4.3.

C.4.21 Proof of Theorem 4.12

Given $\alpha \in \mathbb{S}^{p-1}, \beta \in \mathbb{S}^{q-1}$, expand the square term to have

$$\begin{aligned}
&\left\{4\mathbb{P}\left(\alpha^\top(X_1 - X_2) < 0, \beta^\top(Y_1 - Y_2) < 0\right) - 1\right\}^2 \\
=\ & 16\mathbb{E}\Big[\mathbb{1}(\alpha^\top(X_1 - X_2) < 0, \alpha^\top(X_3 - X_4) < 0) \times \mathbb{1}(\beta^\top(Y_1 - Y_2) < 0, \beta^\top(Y_3 - Y_4) < 0)\Big] \\
& -8\mathbb{E}\left[\mathbb{1}(\alpha^\top(X_1 - X_2) < 0) \times \mathbb{1}(\beta^\top(Y_1 - Y_2) < 0)\right] + 1.
\end{aligned}$$

By applying Lemma 4.0.2, the first term becomes

$$\mathbb{E}\left[\left(2 - \frac{2}{\pi}\mathsf{Ang}\left(X_1 - X_2, X_3 - X_4\right)\right) \cdot \left(2 - \frac{2}{\pi}\mathsf{Ang}\left(Y_1 - Y_2, Y_3 - Y_4\right)\right)\right]$$

and the remainder terms become -1, which yields the expression.

C.4.22 Proof of Theorem 4.13

From Bergsma and Dassios (2014), the univariate τ^* can be written as

$$\begin{aligned}
\tau^* = \ & 4\mathbb{P}\left(X_1 \vee X_2 < X_3 \wedge X_4,\ Y_1 \vee Y_2 < Y_3 \wedge Y_4\right) \\
& + 4\mathbb{P}\left(X_1 \vee X_2 < X_3 \wedge X_4,\ Y_1 \wedge Y_2 > Y_3 \vee Y_4\right) \\
& - 8\mathbb{P}\left(X_1 \vee X_2 < X_3 \wedge X_4,\ Y_1 \vee Y_3 < Y_2 \wedge Y_4\right).
\end{aligned}$$

Notice that

$$\begin{aligned}
&\mathbb{1}(X_1 \vee X_2 < X_3 \wedge X_4) \\
=\ &\mathbb{1}(X_1 < X_2 < X_3 < X_4) + \mathbb{1}(X_2 < X_1 < X_3 < X_4) \\
+\ &\mathbb{1}(X_1 < X_2 < X_4 < X_3) + \mathbb{1}(X_2 < X_1 < X_4 < X_3) \\
=\ &\mathbb{1}(X_1 < X_2)\mathbb{1}(X_2 < X_3)\mathbb{1}(X_3 < X_4) + \mathbb{1}(X_2 < X_1)\mathbb{1}(X_1 < X_3)\mathbb{1}(X_3 < X_4) \\
+\ &\mathbb{1}(X_1 < X_2)\mathbb{1}(X_2 < X_4)\mathbb{1}(X_4 < X_3) + \mathbb{1}(X_2 < X_1)\mathbb{1}(X_1 < X_4)\mathbb{1}(X_4 < X_3).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{1}(Y_1 \vee Y_2 < Y_3 \wedge Y_4) \\
=\ &\mathbb{1}(Y_1 < Y_2)\mathbb{1}(Y_2 < Y_3)\mathbb{1}(Y_3 < Y_4) + \mathbb{1}(Y_2 < Y_1)\mathbb{1}(Y_1 < Y_3)\mathbb{1}(Y_3 < Y_4) \\
+\ &\mathbb{1}(Y_1 < Y_2)\mathbb{1}(Y_2 < Y_4)\mathbb{1}(Y_4 < Y_3) + \mathbb{1}(Y_2 < Y_1)\mathbb{1}(Y_1 < Y_4)\mathbb{1}(Y_4 < Y_3).
\end{aligned}$$

Therefore, the product $I(X_1 \vee X_2 < X_3 \wedge X_4)\mathbb{1}(Y_1 \vee Y_2 < Y_3 \wedge Y_4)$ can be expressed as the linear combination of

$$\mathbb{1}(X_{i_1} < X_{i_2})\mathbb{1}(X_{i_2} < X_{i_3})\mathbb{1}(X_{i_3} < X_{i_4})\mathbb{1}(Y_{j_1} < Y_{j_2})\mathbb{1}(Y_{j_2} < Y_{j_3})\mathbb{1}(Y_{j_3} < Y_{j_4}).$$

Using Lemma C.1.8,

$$\begin{aligned}
&\int_{\mathbb{S}^{p-1}} \mathbb{1}(\alpha^\top X_{i_1} < \alpha^\top X_{i_2})\mathbb{1}(\alpha^\top X_{i_2} < \alpha^\top X_{i_3})\mathbb{1}(\alpha^\top X_{i_3} < \alpha^\top X_{i_4}) d\lambda(\alpha) \\
&= \frac{1}{2} - \frac{1}{4\pi}\left[\mathsf{Ang}\left(U_1, U_2\right) + \mathsf{Ang}\left(U_1, U_3\right) + \mathsf{Ang}\left(U_2, U_3\right)\right],
\end{aligned}$$

where $U_1 = X_{i_1} - X_{i_2}$, $U_2 = X_{i_2} - X_{i_3}$ and $U_3 = X_{i_3} - X_{i_4}$.

Similarly,

$$\begin{aligned}
&\int_{\mathbb{S}^{q-1}} \mathbb{1}(\beta^\top Y_{j_1} < \beta^\top Y_{j_2})\mathbb{1}(\beta^\top Y_{j_2} < \beta^\top Y_{j_3})\mathbb{1}(\beta^\top Y_{j_3} < \beta^\top Y_{j_4}) d\lambda(\beta) \\
&= \frac{1}{2} - \frac{1}{4\pi}\left[\mathsf{Ang}\left(V_1, V_2\right) + \mathsf{Ang}\left(V_1, V_3\right) + \mathsf{Ang}\left(V_2, V_3\right)\right],
\end{aligned}$$

where $V_1 = Y_{j_1} - Y_{j_2}$, $V_2 = Y_{j_2} - Y_{j_3}$ and $V_3 = Y_{j_3} - Y_{j_4}$.

As a result, we have

$$\begin{aligned}
&\int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{q-1}} \mathbb{P}(\alpha^\top X_1 \vee \alpha^\top X_2 < \alpha^\top X_3 \wedge \alpha^\top X_4, \\
&\qquad\qquad \beta^\top Y_1 \vee \beta^\top Y_2 < \beta^\top Y_3 \wedge \beta^\top Y_4) d\lambda(\alpha) d\lambda(\beta) \\
=\ &\mathbb{E}\left[h_p(X_1, X_2, X_3, X_4) h_q(Y_1, Y_2, Y_3, Y_4)\right].
\end{aligned}$$

Applying the same argument to the rest, we can obtain the explicit expression for $\tau^*_{p,q}$ as in Theorem 4.13.

C.4.23 Proof of Theorem C.1

Let us write

$$\begin{aligned}
U^*_{m,n}(Z_{m,n}) &:= U^*_{m,n}(Z_1, \ldots, Z_N) \\
&= N\{U_{m,n}(Z_1, \ldots, Z_N) - \mathbb{E}\left[U_{m,n}(Z_1, \ldots, Z_N)\right]\}
\end{aligned}$$

and denote $U^*_{m,n}(Z_{\varpi(1)}, \ldots, Z_{\varpi(N)})$ by $U^*_{m,n}(Z_\varpi)$. Our goal is to show that for two independent random permutations ϖ, ϖ',

$$\left(U^*_{m,n}(Z_\varpi),\ U^*_{m,n}(Z_{\varpi'})\right) \xrightarrow{d} (T, T'), \tag{C.46}$$

where T, T' are independent and identically distributed with the distribution function $R(t)$. Then the desired result follows by Lemma C.1.4. The proof consists of several pieces and closely follows the proof of the limiting distribution of a two-sample degenerate U-statistic in Chapter 3 of Bhat (1995).

We start with the projection of the two-sample U-statistic via Hoffding's decomposition. Consider the projection of the two-sample degenerate U-statistic based on $Z_{m,n}$:

$$\begin{aligned}
\widehat{U}_{m,n}(Z_{m,n}) = \ &\frac{r(r-1)}{m(m-1)} \sum_{1 \le i_1 < i_2 \le m} g^*_{2,0}(Z_{i_1}, Z_{i_2}) \\
&+ \frac{r(r-1)}{n(n-1)} \sum_{1 \le j_1 < j_2 \le n} g^*_{0,2}(Z_{j_1+m}, Z_{j_2+m}) + \frac{r^2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} g^*_{1,1}(Z_i, Z_{j+m}).
\end{aligned}$$

Then it can be seen that

$$\begin{aligned}
&\mathbb{E}[(U_{m,n}(Z_{m,n}) - \widehat{U}_{m,n}(Z_{m,n})] = 0 \text{ and} \\
&\mathrm{Var}[U_{m,n}(Z_{m,n}) - \widehat{U}_{m,n}(Z_{m,n})] = O(N^{-3}),
\end{aligned}$$

which implies

$$N(U_{m,n}(Z_{m,n}) - \theta) = N(\widehat{U}_{m,n}(Z_{m,n}) - \theta) + o_{\mathbb{P}}(1). \tag{C.47}$$

Under the finite second moment of the kernel g, we may have the decompositions

$$g^*_{2,0}(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y),$$

$$g^*_{0,2}(x,y) = \sum_{i=1}^{\infty} \gamma_i \psi_i(x)\psi_i(y),$$

$$g^*_{1,1}(x,y) = \sum_{i=1}^{\infty} \alpha_i \phi^*_i(x)\psi^*_i(y),$$

where $\{\phi_i(\cdot)\}$, $\{\psi_i(\cdot)\}$, $\{\phi^*(\cdot), \psi^*(\cdot)\}$ are orthonormal eigenfunctions and the corresponding eigenvalues $\{\lambda_i\}, \{\gamma_i\}, \{\alpha_i\}$, associated with $g^*_{2,0}, g^*_{0,2}$ and $g^*_{1,1}$, respectively (see e.g. Bhat, 1995, for details). From the given conditions of the theorem, the eigenvalues and the eigenfunctions are related as follows:

$$\phi_i(z) = \psi_i(z) = \phi^*_i(z) = \psi^*_i(z),$$

$$\gamma_i = \lambda_i \quad \text{and} \quad \alpha_i = \frac{1-r}{r}\lambda_i.$$

Therefore,

$$\begin{aligned} N\widehat{U}_{m,n}(Z_{m,n}) = \; & \widehat{a}_1 \left[\frac{1}{m} \sum_{1 \le i_1 \ne i_2 \le m} \sum_{i=1}^{\infty} \lambda_i \phi_i(Z_{i_1})\phi_i(Z_{i_2}) \right] \\ & + \widehat{a}_2 \left[\frac{1}{n} \sum_{1 \le j_1 \ne j_2 \le n} \sum_{j=1}^{\infty} \lambda_j \phi_j(Z_{j_1+m})\phi_j(Z_{j_2+m}) \right] \\ & + \widehat{a}_3 \left[\frac{1}{\sqrt{mn}} \sum_{i_1=1}^{m} \sum_{j_1=1}^{n} \sum_{k=1}^{\infty} \lambda_k \phi_k(Z_{i_1})\phi_k(Z_{j_1+m}) \right] \\ = \; & \widehat{a}_1 T_m + \widehat{a}_2 T'_n + \widehat{a}_3 T''_{mn}, \end{aligned}$$

where

$$\widehat{a}_1 = \frac{r(r-1)}{2}\frac{N}{m-1}, \quad \widehat{a}_2 = \frac{r(r-1)}{2}\frac{N}{n-1} \quad \text{and} \quad \widehat{a}_3 = -r(r-1)\frac{N}{\sqrt{mn}}.$$

Denote the centered and scaled projection of the U-statistic by

$$\widetilde{U}_{m,n} := N(\widehat{U}_{m,n}(Z_{\varpi}) - \theta) \quad \text{and} \quad \widetilde{U}'_{m,n} := N(\widehat{U}_{m,n}(Z_{\varpi'}) - \theta).$$

Then due to (C.47),

$$\left(U^*_{m,n}(Z_{\varpi}),\ U^*_{m,n}(Z_{\varpi'})\right) = \left(\widetilde{U}_{m,n}(Z_{\varpi}),\ \widetilde{U}'_{m,n}(Z_{\varpi'})\right) + o_{\mathbb{P}}(1).$$

Therefore it suffices to show

$$\left(\widetilde{U}_{m,n},\ \widetilde{U}'_{m,n}\right) \xrightarrow{d} (T, T')$$

to complete the main proof. Having this goal in mind, we start with a truncation of the degenerate U-statistic.

• **Truncation of the U-statistics.**

Now, define a truncated version of $N(\widehat{U}_{m,n}(Z_{m,n}) - \theta)$ by

$$\begin{aligned} N(\widehat{U}_{m,n,K}(Z_{m,n}) - \theta) = \ & \widehat{a}_1 \left[\frac{1}{m} \sum_{1 \le i_1 \ne i_2 \le m} \sum_{i=1}^{K} \lambda_i \phi_i(Z_{i_1}) \phi_i(Z_{i_2}) \right] \\ & + \widehat{a}_2 \left[\frac{1}{n} \sum_{1 \le j_1 \ne j_2 \le n} \sum_{j=1}^{K} \lambda_j \phi_j(Z_{j_1+m}) \phi_j(Z_{j_2+m}) \right] \\ & + \widehat{a}_3 \left[\frac{1}{\sqrt{mn}} \sum_{i_1=1}^{m} \sum_{j_1=1}^{n} \sum_{k=1}^{K} \lambda_k \phi_k(Z_{i_1}) \phi_k(Z_{j_1+m}) \right] \\ = \ & \widehat{a}_1 T_{mK} + \widehat{a}_2 T'_{nK} + \widehat{a}_3 T''_{mnK}. \end{aligned} \tag{C.48}$$

Write

$$\begin{aligned} & \widehat{a}_1 T_{mK} + \widehat{a}_2 T'_{nK} + \widehat{a}_3 T''_{mnK} \\ = \ & \widehat{a}_1 \left[\sum_{k=1}^{K} \lambda_k \left(W^2_{km} - V_{km} \right) \right] + \widehat{a}_2 \left[\sum_{k=1}^{K} \lambda_k \left(W'^2_{kn} - V'_{kn} \right) \right] + \widehat{a}_3 \left[\sum_{k=1}^{K} \lambda_k W_{km} W'_{kn} \right] \\ = \ & \frac{r(r-1)}{2} \left\{ \sum_{k=1}^{K} \lambda_k \left(\sqrt{\frac{N}{m}} W_{km} - \sqrt{\frac{N}{n}} W'_{kn} \right)^2 - \sum_{k=1}^{K} \lambda_k \left(\frac{N}{m} V_{km} + \frac{N}{n} V'_{kn} \right) \right\}, \end{aligned}$$

where

$$W_{km} = \frac{1}{\sqrt{m}} \sum_{i_1=1}^{m} \phi_k(Z_{i_1}), \quad W'_{kn} = \frac{1}{\sqrt{n}} \sum_{j_1=1}^{n} \phi_k(Z_{j_1+m}),$$

$$V_{km} = \frac{1}{m} \sum_{i_1=1}^{m} \phi_k^2(Z_{i_1}), \quad V'_{kn} = \frac{1}{n} \sum_{j_1=1}^{n} \phi_k^2(Z_{j_1+m}),$$

for $k = 1, \ldots, K$.

By strong law of large numbers,

$$V_{mn}^{*\top} := (V_{1m}, \ldots, V_{Km}, V'_{1n}, \ldots, V'_{Kn})^\top \xrightarrow{a.s.} V^{*\top} = (V_1, \ldots, V_K, V'_1, \ldots, V'_K)^\top$$

and by the assumption that $m/N \to \vartheta_X$, $n/N \to \vartheta_Y$,

$$\begin{aligned}
& N(\widehat{U}_{m,n,K} - \theta) \\
= \; & \frac{r(r-1)}{2} \left\{ \sum_{k=1}^{K} \lambda_k \left(\sqrt{\frac{N}{m}} W_{km} - \frac{r(r-1)}{2} \sqrt{\frac{N}{n}} W'_{kn} \right)^2 - \frac{1}{\vartheta_X \vartheta_Y} \sum_{k=1}^{K} \lambda_k \right\} \\
& + o_{\mathbb{P}}(1) \\
= \; & \frac{r(r-1)}{2} \left\{ N \sum_{k=1}^{K} \lambda_k \left(\frac{1}{m} \sum_{i=1}^{m} \phi_k(Z_i) - \frac{1}{n} \sum_{j=1}^{n} \phi_k(Z_{j+m}) \right)^2 - \frac{1}{\vartheta_X \vartheta_Y} \sum_{k=1}^{K} \lambda_k \right\} \\
& + o_{\mathbb{P}}(1) \\
= \; & \frac{r(r-1)}{2} \left\{ N \sum_{k=1}^{K} \lambda_k \left(\sum_{i=1}^{N} \epsilon_i \phi_k(Z_i) \right)^2 - \frac{1}{\vartheta_X \vartheta_Y} \sum_{k=1}^{K} \lambda_k \right\} + o_{\mathbb{P}}(1)
\end{aligned}$$

where

$$(\epsilon_1, \ldots, \epsilon_m, \epsilon_{m+1}, \ldots, \epsilon_{m+n}) = (\underbrace{m^{-1}, \ldots, m^{-1}}_{m \text{ terms}}, \underbrace{-n^{-1}, \ldots, -n^{-1}}_{n \text{ terms}}).$$

• **Proving independence of the truncated U-statistics.**

Consider the truncated permutation statistics

$$\widetilde{U}_{m,n,K} := N(\widehat{U}_{m,n,K}(Z_\varpi) - \theta)$$

$$= \frac{r(r-1)}{2}\left\{ N\sum_{k=1}^{K}\lambda_k\left(\sum_{i=1}^{N}\epsilon_{\varpi(i)}\phi_k(Z_i)\right)^2 - \frac{1}{\vartheta_X\vartheta_Y}\sum_{k=1}^{K}\lambda_k\right\} + o_{\mathbb{P}}(1)$$

$$\widetilde{U}'_{m,n,K} := N(\widehat{U}_{m,n,K}(Z_{\varpi'}) - \theta)$$

$$= \frac{r(r-1)}{2}\left\{ N\sum_{k=1}^{K}\lambda_k\left(\sum_{i=1}^{N}\epsilon_{\varpi'(i)}\phi_k(Z_i)\right)^2 - \frac{1}{\vartheta_X\vartheta_Y}\sum_{k=1}^{K}\lambda_k\right\} + o_{\mathbb{P}}(1).$$

Note that $\epsilon_{\varpi(i)}$ and $\epsilon_{\varpi'(i)}$ are independent random variables by the assumption having either $1/m$ or $-1/n$ with m/N and n/N probabilities; hence

$$\mathrm{Cov}\left(\epsilon_{\varpi(i)}\phi_k(Z_i), \epsilon_{\varpi'(i)}\phi_k(Z_i)\right) = \mathbb{E}\left[\epsilon_{\varpi(i)}\right]\mathbb{E}\left[\epsilon_{\varpi'(i)}\right]\mathbb{E}\left[\phi_k^2(Z_i)\right] = 0.$$

By the Cramér-Wold device and the Lindeberg condition, we see that

$$\sqrt{N}\left(\sum_{i=1}^{N}\epsilon_{\varpi(i)}\phi_1(Z_i), \ldots, \sum_{i=1}^{N}\epsilon_{\varpi(i)}\phi_K(Z_i), \right.$$
$$\left. \sum_{i=1}^{N}\epsilon_{\varpi'(i)}\phi_1(Z_i), \ldots, \sum_{i=1}^{N}\epsilon_{\varpi'(i)}\phi_K(Z_i)\right)^\top$$
$$\xrightarrow{d} N(0, {\vartheta_X}^{-1}{\vartheta_Y}^{-1}I_{2K}).$$

Thus the components of the vector are asymptotically independent to each other. Then apply the continuous mapping theorem together with Slutsky's theorem to have

$$(\widetilde{U}_{m,n,K}, \widetilde{U}'_{m,n,K}) \xrightarrow{d} (T_K, T'_K) \tag{C.49}$$

where T_K and T'_K are independent and have the same distribution as

$$\frac{r(r-1)}{2\vartheta_X\vartheta_Y}\sum_{k=1}^{K}\lambda_k(\xi_k^2 - 1),$$

where $\xi_k \overset{i.i.d.}{\sim} N(0,1)$.

• **Bounding the difference between characteristic functions.**

We will use the characteristic functions to show

$$\left(\widetilde{U}_{m,n},\ \widetilde{U}'_{m,n}\right) \xrightarrow{d} (T, T').$$

More specifically, we will show that for any $x, y \in \mathbb{R}$ and any $\epsilon > 0$ and sufficiently large N,

$$\left| \mathbb{E}\left[e^{i(x\widetilde{U}_{m,n} + y\widetilde{U}'_{m,n})} \right] - \mathbb{E}\left[e^{i(xT + yT')} \right] \right| \leq \ (I) + (II) + (III) \ < \epsilon$$

where

$$\begin{aligned}
(I) &= \left| \mathbb{E}\left[e^{i(x\widetilde{U}_{m,n} + y\widetilde{U}'_{m,n})} \right] - \mathbb{E}\left[e^{i(x\widetilde{U}_{m,n,K} + y\widetilde{U}'_{m,n,K})} \right] \right|, \\
(II) &= \left| \mathbb{E}\left[e^{i(x\widetilde{U}_{m,n,K} + y\widetilde{U}'_{m,n,K})} \right] - \mathbb{E}\left[e^{i(xT_K + yT'_K)} \right] \right|, \\
(III) &= \left| \mathbb{E}\left[e^{i(xT_K + yT'_K)} \right] - \mathbb{E}\left[e^{i(xT + yT')} \right] \right|.
\end{aligned}$$

We bound these terms in sequence.

1. Bounding (I).

Based on $|e^{iz}| = 1$ and $|e^{iz} - 1| \leq |z|$, we bound (I) by

$$\begin{aligned}
(I) \ &= \ \left| \mathbb{E}\left[e^{i(x\widetilde{U}_{m,n} + y\widetilde{U}'_{m,n})} \right] - \mathbb{E}\left[e^{i(x\widetilde{U}_{m,n,K} + y\widetilde{U}'_{m,n,K}))} \right] \right| \\
&\leq \ |x| \left[\mathbb{E}\left(\widetilde{U}_{m,n,K} - \widetilde{U}_{m,n} \right)^2 \right]^{1/2} + |y| \left[\mathbb{E}\left(\widetilde{U}'_{m,n,K} - \widetilde{U}'_{m,n} \right)^2 \right]^{1/2} \\
&\leq \ (|x| + |y|) \left\{ \frac{r(r-1)}{2\widehat{\vartheta}_1} \left(2 \sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} + \frac{r(r-1)}{2\widehat{\vartheta}_2} \left(2 \sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} \right. \\
&\qquad \left. - \frac{r(r-1)}{\sqrt{\widehat{\vartheta}_1 \widehat{\vartheta}_2}} \left(\sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} \right\} \\
&= \ (|x| + |y|) \frac{r(r-1)}{\sqrt{2}} \left(\frac{1}{\sqrt{\widehat{\vartheta}_1}} - \frac{1}{\sqrt{\widehat{\vartheta}_2}} \right)^2 \left(\sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} \\
&\leq \ (|x| + |y|) \frac{r(r-1)}{\sqrt{2}\widehat{\vartheta}_1 \widehat{\vartheta}_2} \left(\sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2}
\end{aligned}$$

where $\widehat{\vartheta}_1 = m/N$ and $\widehat{\vartheta}_2 = n/N$.

Now, for fixed x and y and any given $\epsilon > 0$, we choose K large enough to bound

$$(|x| + |y|) \frac{r(r-1)}{\sqrt{2}\vartheta_X \vartheta_Y} \left(\sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} < \frac{\epsilon}{3}. \tag{C.50}$$

Since $\widehat{\vartheta}_1 \to \vartheta_X$ and $\widehat{\vartheta}_2 \to \vartheta_Y$ as $N \to \infty$, we have

$$(I) \;\leq\; (|x| + |y|)\, \frac{r(r-1)}{\sqrt{2}\widehat{\vartheta}_1 \widehat{\vartheta}_2} \left(\sum_{k=K+1}^{\infty} \lambda_k^2 \right)^{1/2} < \frac{\epsilon}{3},$$

for all sufficiently large N.

2. Bounding (II).

From the result established in (C.49), we have

$$(II) \;=\; \left| \mathbb{E}\left[e^{i(x\widetilde{U}_{m,n,K} + y\widetilde{U}'_{m,n,K})} \right] - \mathbb{E}\left[e^{i(xT_K + yT'_K)} \right] \right| < \frac{\epsilon}{3}$$

for all sufficiently large N.

3. Bounding (III).

From Chapter 3 of Bhat (1995) with the conditions given on the kernel, the asymptotic distribution of a degenerate U-statistic converges to

$$\begin{aligned} N\left(U_{m,n} - \theta\right) \;\xrightarrow{d}\; & \frac{r(r-1)}{2\vartheta_X} \sum_{k=1}^{\infty} \lambda_k(\xi_k^2 - 1) + \frac{r(r-1)}{2\vartheta_Y} \sum_{k=1}^{\infty} \lambda_k(\xi_k'^2 - 1) \\ & - \frac{r(r-1)}{\sqrt{\vartheta_X \vartheta_Y}} \sum_{k=1}^{\infty} \lambda_k \xi_k \xi_k' \end{aligned} \tag{C.51}$$

where $\{\xi_k\}$ and $\{\xi_k'\}$ are independent standard normal random variables and $\{\lambda_k\}$ are eigenvalues associated with the kernel. Note that the right-side of (C.51) can be re-written as

$$\frac{r(r-1)}{2\vartheta_X \vartheta_Y} \sum_{k=1}^{\infty} \lambda_k \left[(\sqrt{\vartheta_Y}\xi_k - \sqrt{\vartheta_X}\xi_k')^2 - 1 \right],$$

where $\sqrt{\vartheta_Y}\xi_k - \sqrt{\vartheta_X}\xi_k' \sim N(0,1)$. Therefore, T, T' are identically distributed as

$$\frac{r(r-1)}{2\vartheta_X \vartheta_Y} \sum_{k=1}^{\infty} \lambda_k(\xi_k^2 - 1).$$

Recall that T_K, T'_K have the same distribution as

$$\frac{r(r-1)}{2\vartheta_X \vartheta_Y} \sum_{k=1}^{K} \lambda_k(\xi_k^2 - 1).$$

Consequently,

$$\begin{aligned}
&\left|\mathbb{E}\left[e^{i(xT_K+yT'_K)}\right]-\mathbb{E}\left[e^{i(xT+yT')}\right]\right| \\
\leq\ & |x|\left[\mathbb{E}\left(T_K-T\right)^2\right]^{1/2}+|y|\left[\mathbb{E}\left(T'_K-T'\right)^2\right]^{1/2} \\
\leq\ & (|x|+|y|)\,\frac{r(r-1)}{\sqrt{2}\vartheta_X\vartheta_Y}\left(\sum_{k=K+1}^{\infty}\lambda_k^2\right)^{1/2}<\frac{\epsilon}{3},
\end{aligned}$$

with the same choice of x, y, ϵ, K in (C.50).

• **Combining the bounds.**

From the previous results, we conclude that for any $x, y \in \mathbb{R}$ and any $\epsilon > 0$ with sufficiently large N,

$$\left|\mathbb{E}\left[e^{i(x\widetilde{U}_{m,n}+y\widetilde{U}'_{m,n})}\right]-\mathbb{E}\left[e^{i(xT+yT')}\right]\right|<\epsilon,$$

and therefore

$$\left(\widetilde{U}_{m,n},\ \widetilde{U}'_{m,n}\right)\xrightarrow{d}(T,T').$$

This completes the proof.

C.5 Additional Results

This section presents some details omitted in the main text and also additional results. In particular,

- Appendix C.5.1 verifies the condition (4.16) in the main text.
- Appendix C.5.2 extends Lemma 4.0.2 to the integration involving four indicator functions.
- Appendix C.5.3 provides details on Remark 4.8 and shows the asymptotic equivalence between projection-averaging and spatial-sign statistics.
- Appendix C.5.4 collects some variants of the CvM-statistic.
- Appendix C.5.5 describes the power functions of the tests in the HDLSS setting.
- Appendix C.5.6 proves that the angular distance is a metric of negative-type.

- Appendix C.5.7 provides details on Remark 4.6.

- Appendix C.5.8 collects further applications of projection-averaging.

C.5.1 Verification of (4.16) in the main text

First we state the distributional assumptions made in Bai and Saranadasa (1996) and Chen and Qin (2010):

$$X = \Gamma_X V_X + \mu_X \quad \text{and} \quad Y = \Gamma_Y V_Y + \mu_Y, \tag{C.52}$$

where V_X and V_Y are independent random vectors in $\mathbb{R}^u$ for some $u \geq d$ such that $\mathbb{E}(V_X) = \mathbb{E}(V_Y) = 0$ and $\mathrm{Var}(V_X) = \mathrm{Var}(V_Y) = I_u$, the $u \times u$ identity matrix. Γ_X and Γ_Y are non-random $d \times u$ matrices such that $\Sigma_X = \Gamma_X \Gamma_X^\top$ and $\Sigma_Y = \Gamma_Y \Gamma_Y^\top$ are positive definite and μ_X and μ_Y are non-random d-dimensional vectors. Write $V_X = (V_{X,1}, \ldots, V_{X,m})$ and $V_Y = (V_{Y,1}, \ldots, V_{Y,m})$. Assume that $\mathbb{E}(V_{X,i}^4) = \mathbb{E}(V_{Y,i}) = 3 + \Delta < \infty$ for $i = 1, \ldots, m$ where Δ is the difference between the fourth moment of $V_{X,i}$ and $N(0,1)$. In addition assume that

$$\mathbb{E}(V_{X,l_1}^{\alpha_1} V_{X,l_2}^{\alpha_2} \cdots V_{X,l_q}^{\alpha_q}) = \prod_{i=1}^{q} \mathbb{E}(V_{X,l_i}^{\alpha_i}) \quad \text{and}$$

$$\mathbb{E}(V_{Y,l_1}^{\alpha_1} V_{Y,l_2}^{\alpha_2} \cdots V_{Y,l_q}^{\alpha_q}) = \prod_{i=1}^{q} \mathbb{E}(V_{Y,l_i}^{\alpha_i})$$

for a positive integer q such that $\sum_{l=1}^{q} \alpha_l \leq 8$ and $l_1 \neq l_2 \neq \cdots \neq l_q$.

Our goal here is to show that $\mathrm{Var}(\|Z_1 - Z_2\|^2) = O(d)$ and $\mathrm{Var}\{(Z_1 - Z_3)^\top (Z_2 - Z_3)\} = O(d)$ are implied by

$$(\mu_X - \mu_Y)^\top (\Sigma_X + \Sigma_Y)(\mu_X - \mu_Y) = O(d) \quad \text{and}$$

$$\mathrm{tr}\{(\Sigma_X + \Sigma_Y)^2\} = O(d),$$

where Z_1, Z_2, Z_3 are independent and each Z_i is identically distributed as either X or Y in the model (C.52). First let us focus on $\mathrm{Var}(\|Z_1 - Z_2\|^2)$. Denote $\overline{Z}_1 = Z_1 - \mathbb{E}(Z_1)$, $\overline{Z}_2 = Z_2 - \mathbb{E}(Z_2)$ and $\delta_{12} = \mathbb{E}(Z_1) - \mathbb{E}(Z_2)$. Based on the basic inequality that

$$\mathrm{Var}\Big(\sum_{i=1}^{k} X_i\Big) \leq k \sum_{i=1}^{k} \mathrm{Var}(X_i) \quad \text{for any } k \geq 1,$$

we have

$$\begin{aligned}\mathrm{Var}(\|Z_1 - Z_2\|^2) &= \mathrm{Var}\{(\overline{Z}_1 - \overline{Z}_2)^\top(\overline{Z}_1 - \overline{Z}_2) + 2\delta_{12}^\top(\overline{Z}_1 - \overline{Z}_2)\} \\ &\leq 8\mathrm{Var}(\overline{Z}_1^\top \overline{Z}_1) + 8\mathrm{Var}(\overline{Z}_2^\top \overline{Z}_2) + 16\mathrm{Var}(\overline{Z}_1^\top \overline{Z}_2) \\ &\quad + 8\delta_{12}^\top \mathrm{Var}(\overline{Z}_1 - \overline{Z}_2)\delta_{12}.\end{aligned}$$

Now using Proposition A.1 of **?**, we have that

$$\begin{aligned}\mathrm{Var}(\overline{Z}_1^\top \overline{Z}_1) &\leq (2 + \Delta)\mathrm{tr}(\Sigma_{Z_1}^2) \quad \text{and} \\ \mathrm{Var}(\overline{Z}_2^\top \overline{Z}_2) &\leq (2 + \Delta)\mathrm{tr}(\Sigma_{Z_2}^2),\end{aligned}$$

where $\Sigma_{Z_i} = \mathrm{Var}(Z_i)$ for $i = 1, 2$. Additionally we know that $\mathrm{Var}(\overline{Z}_1^\top \overline{Z}_2) \leq \mathbb{E}\{(\overline{Z}_1^\top \overline{Z}_2)^2\} = \mathrm{tr}(\Sigma_{Z_1}\Sigma_{Z_2})$. Combining the results,

$$\mathrm{Var}(\|Z_1 - Z_2\|^2) \lesssim \mathrm{tr}\{(\Sigma_X + \Sigma_Y)^2\} + (\mu_X - \mu_Y)^\top(\Sigma_X + \Sigma_Y)(\mu_X - \mu_Y).$$

Hence $\mathrm{Var}(\|Z_1 - Z_2\|^2) = O(d)$ under the model assumption (4.16).

Next moving onto $\mathrm{Var}\{(Z_1 - Z_3)^\top(Z_2 - Z_3)\}$, write $\overline{Z}_3 = Z_3 - \mathbb{E}(Z_3)$, $\delta_{13} = \mathbb{E}(Z_1) - \mathbb{E}(Z_3)$ and $\delta_{23} = \mathbb{E}(Z_2) - \mathbb{E}(Z_3)$. Then it can be shown that

$$\begin{aligned}&\mathrm{Var}\{(Z_1 - Z_3)^\top(Z_2 - Z_3)\} \\ \leq\ & 12\mathrm{Var}(\overline{Z}_1^\top \overline{Z}_2) + 12\mathrm{Var}(\overline{Z}_1^\top \overline{Z}_3) + 12\mathrm{Var}(\overline{Z}_3^\top \overline{Z}_2) + 12\mathrm{Var}(\overline{Z}_3^\top \overline{Z}_3) \\ &+ 3\delta_{13}^\top \mathrm{Var}(\overline{Z}_2 - \overline{Z}_3)\delta_{13} + 3\delta_{23}^\top \mathrm{Var}(\overline{Z}_1 - \overline{Z}_3)\delta_{23}.\end{aligned}$$

Now similarly as before,

$$\mathrm{Var}\{(Z_1 - Z_3)^\top(Z_2 - Z_3)\} \lesssim \mathrm{tr}\{(\Sigma_X + \Sigma_Y)^2\} + (\mu_X - \mu_Y)^\top(\Sigma_X + \Sigma_Y)(\mu_X - \mu_Y).$$

Hence $\mathrm{Var}\{(Z_1 - Z_3)^\top(Z_2 - Z_3)\} = O(d)$ under the model assumption (4.16).

C.5.2 Generalization of Lemma 4.0.2 and Lemma C.1.8

Here we provide the explicit formula for the integration involving four indicator functions, which extends Lemma 4.0.2 in the main text and Lemma C.1.8 in this supplementary material.

Lemma C.1.30 (Extension of Escanciano (2006)). *For arbitrary non-zero vectors $U_1, U_2, U_3, U_4 \in \mathbb{R}^d$, let us denote $\varrho_{ij} = U_i U_j / \{\|U_i\|\|U_j\|\}$ for $i, j \in \{1, 2, 3, 4\}$. Then*

$$\int_{\mathbb{S}^{d-1}} \prod_{i=1}^{4} \mathbb{1}(\beta^\top U_i \le 0) d\lambda(\beta) = \frac{7}{16} + \frac{1}{8\pi} \sum_{i=1}^{3} \sum_{j=i+1}^{4} \mathsf{Ang}(U_i, U_j) + Q, \tag{C.53}$$

where

$$Q = \frac{1}{4\pi^2} \sum_{\ell=1}^{4} \int_0^1 \frac{\varrho_{1\ell}}{(1 - \varrho_{1\ell}^2 u^2)^{1/2}} \arcsin\left\{ \frac{\gamma_{1,\ell}(u)}{\gamma_{2,\ell}(u)\gamma_{3,\ell}(u)} \right\} du,$$

with

$$\begin{aligned}
\gamma_{1,2} &= \varrho_{34} - \varrho_{23}\varrho_{24} - [\varrho_{13}\varrho_{14} + \varrho_{12}(\varrho_{12}\varrho_{34} - \varrho_{14}\varrho_{23} - \varrho_{13}\varrho_{24})]u^2, \\
\gamma_{1,3} &= \varrho_{24} - \varrho_{23}\varrho_{34} - [\varrho_{12}\varrho_{14} + \varrho_{13}(\varrho_{13}\varrho_{24} - \varrho_{14}\varrho_{23} - \varrho_{12}\varrho_{34})]u^2, \\
\gamma_{1,4} &= \varrho_{23} - \varrho_{24}\varrho_{34} - [\varrho_{12}\varrho_{13} + \varrho_{14}(\varrho_{14}\varrho_{23} - \varrho_{13}\varrho_{24} - \varrho_{12}\varrho_{34})]u^2, \\
\gamma_{2,2} &= \gamma_{2,3} = [1 - \varrho_{23}^2 - (\varrho_{12}^2 + \varrho_{13}^2 - 2\varrho_{12}\varrho_{13}\varrho_{23})u^2]^{1/2}, \\
\gamma_{3,2} &= \gamma_{2,4} = [1 - \varrho_{24}^2 - (\varrho_{12}^2 + \varrho_{14}^2 - 2\varrho_{12}\varrho_{14}\varrho_{24})u^2]^{1/2}, \\
\gamma_{3,3} &= \gamma_{3,4} = [1 - \varrho_{34}^2 - (\varrho_{13}^2 + \varrho_{14}^2 - 2\varrho_{13}\varrho_{14}\varrho_{34})u^2]^{1/2}.
\end{aligned}$$

Proof. To prove the results, we apply the same argument used in Section C.4.2. Let $\mathcal{Z}$ have a multivariate normal distribution with zero mean vector and identity covariance matrix. Then as in Section C.4.2, we have

$$\int_{\mathbb{S}^{d-1}} \prod_{i=1}^{4} \mathbb{1}(\beta^\top U_i \le 0) d\lambda(\beta) = \mathbb{E}_{\mathcal{Z}}\left[\prod_{i=1}^{4} \mathbb{1}(\mathcal{Z}^\top U_i \le 0) \right]. \tag{C.54}$$

Since $(\mathcal{Z}^\top U_1, \mathcal{Z}^\top U_2, \mathcal{Z}^\top U_3, \mathcal{Z}^\top U_4)^\top$ has a multivariate normal distribution with zero mean vector and correlation matrix $[\varrho_{ij}]_{4\times 4}$ with

$$\varrho_{ij} = \frac{U_i^\top U_j}{\|U_i\|\|U_j\|},$$

the right-hand side of (C.54) can be computed based on orthant probabilities for normal distributions (e.g. Childs, 1967; Xu et al., 2013). This completes the proof. □

Remark C.3. *Although the explicit formula given in Lemma C.1.30 looks complicated, it reduces the integral over $\mathbb{S}^{d-1}$ to a more tractable single integral over the unit interval. Hence it would help significantly improve computational time and efficiency in practical applications.*

Remark C.4. *Childs (1967) also provided expressions for higher order integrations. Using the same argument as before, it is possible to further generalize Lemma C.1.30.*

C.5.3 Asymptotic Equivalence between Projection-Averaging and Spatial-Sign Statistics

In this section, we provide details on Remark 4.8 in the main text. Building on U-statistics, the multivariate one-sample sign test statistic and the two-sample WMW test statistic via projection-averaging can be defined as

$$U_{\text{Sign-Proj}} = \frac{1}{(m)_2} \sum_{i,j=1}^{m,\neq} h_{\text{Sign-Proj}}(X_i, X_j) \quad \text{and}$$

$$U_{\text{WMW-Proj}} = \frac{1}{(m)_2(n)_2} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j_1,j_2=1}^{n,\neq} h_{\text{WMW-Proj}}(X_{i_1}, X_{i_2}; Y_{j_1}, Y_{j_2}),$$

where

$$h_{\text{Sign-Proj}}(x, y) = \frac{1}{4} - \frac{1}{2\pi}\mathsf{Ang}(x, y) \quad \text{and}$$

$$h_{\text{WMW-Proj}}(x_1, x_2; y_1, y_2) = \frac{1}{4} - \frac{1}{2\pi}\mathsf{Ang}(x_1 - y_1, x_2 - y_2).$$

On the other hand, the multivariate one-sample sign test statistic and two-sample WMW test statistic based on the spatial sign are

$$U_{\text{Sign-SS}} = \frac{1}{(m)_2} \sum_{i,j=1}^{m,\neq} \frac{X_i^\top X_j}{\|X_i\|\|X_j\|},$$

$$U_{\text{WMW-SS}} = \frac{1}{(m)_2(n)_2} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j_1,j_2=1}^{n,\neq} \frac{(X_{i_1} - Y_{j_1})^\top (X_{i_2} - Y_{j_2})}{\|X_{i_1} - Y_{j_1}\|\|X_{i_2} - Y_{j_2}\|}.$$

We provide the following proposition for the one-sample case where we prove the asymptotic equivalence between $U_{\text{Sign-Proj}}$ and $U_{\text{Sign-SS}}$.

Proposition C.5. *Suppose that* $Var[X_1^\top X_2] = O(d)$ *and* $Var[\|X_1\|^2] = O(d)$. *Let us write and assume that*

$$\eta_{X,d} = \frac{\|\mu_X\|^2}{\|\mu_X\|^2 + \mathrm{tr}(\Sigma_X)} \to \eta_X \in [0,1),$$
$$\delta_{X,d} = \frac{1}{4} - \frac{1}{2\pi}\arccos(\eta_{X,d}) - \frac{\eta_{X,d}}{2\pi(1-\eta_{X,d}^2)^{1/2}}.$$

Then under the HDLSS setting,

$$U_{\text{Sign-Proj}} = \delta_{X,d} + \frac{1}{2\pi(1-\eta_{X,d}^2)^{1/2}} U_{\text{Sign-SS}} + O_{\mathbb{P}}(d^{-1}).$$

When $\mu_X = 0$, *the expression can be simplified as*

$$U_{\text{Sign-Proj}} = \frac{1}{\sqrt{2\pi}} U_{\text{Sign-SS}} + O_{\mathbb{P}}(d^{-1}).$$

Proof. Similarly as in Section C.4.16, we use the Taylor expansion and the weak law of large numbers to obtain

$$\frac{X_1^\top X_2}{\|X_1\|\|X_2\|} = \eta_{X,d} + O_{\mathbb{P}}(d^{-1/2}).$$

Next applying the second order Taylor expansion of $f(x) = \arccos(x)$ around $f(\eta_{X,d})$ yields

$$\begin{aligned} &\arccos\left\{\frac{X_1^\top X_2}{\|X_1\|\|X_2\|}\right\} \\ =\ &\arccos(\eta_{X,d}) - \frac{1}{(1-\eta_{X,d}^2)^{1/2}}\left(\frac{X_1^\top X_2}{\|X_1\|\|X_2\|} - \eta_{X,d}\right) + O_{\mathbb{P}}(d^{-1}). \end{aligned}$$

We finish the proof by plugging this approximation into $U_{\text{Sign-Proj}}$. □

For the two-sample case, we present the following result.

Proposition C.6. *Suppose that* $Var[(X_1 - Y_1)^\top (X_2 - Y_2)] = O(d)$, $Var[\|X_1 - Y_1\|^2] = O(d)$. *Let us write and assume that*

$$\eta_{XY,d} = \frac{\|\mu_X - \mu_Y\|^2}{\|\mu_X - \mu_Y\|^2 + \mathrm{tr}(\Sigma_X) + \mathrm{tr}(\Sigma_Y)} \to \eta_{XY} \in [0,1).$$
$$\delta_{XY,d} = \frac{1}{4} - \frac{1}{2\pi}\arccos(\eta_{XY,d}) - \frac{\eta_{XY,d}}{2\pi(1-\eta_{XY,d}^2)^{1/2}}.$$

Then under the HDLSS setting,

$$U_{\text{WMW-Proj}} = \delta_{XY,d} + \frac{1}{2\pi(1-\eta_{XY,d}^2)^{1/2}} U_{\text{WMW-SS}} + O_{\mathbb{P}}(d^{-1}).$$

When $\mu_X = \mu_Y$, the expression can be simplified as

$$U_{\text{WMW-Proj}} = \frac{1}{\sqrt{2\pi}} U_{\text{WMW-SS}} + O_{\mathbb{P}}(d^{-1}).$$

The proof of this result is similar to that of Proposition C.5; hence omitted.

C.5.4 Some variants

In this section, we provide several variants of the proposed test statistics. We first present the linear-type test statistic, which is computationally more efficient than U_{CvM}, defined as follows:

$$\begin{aligned} L_{\text{CvM}} \;=\; \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\Big[&h_{\text{CvM}}(X_{2i-1}, X_{2i}; Y_{2i-1}, Y_{2i}) \\ &+ h_{\text{CvM}}(X_{2i}, X_{2i-1}; Y_{2i}, Y_{2i-i})\Big], \end{aligned} \tag{C.55}$$

where $M = \lfloor n/2 \rfloor$ and $m = n$ for simplicity. While L_{CvM} is also an unbiased estimator of W_d^2 and can be computed in linear time, the test based on L_{CvM} is sub-optimal in terms of minimax power. This illustrates a trade-off between computational complexity and statistical power. In detail, we show that the oracle test based on L_{CvM} can have full power only against alternatives shrinking slower than $N^{-1/4}$ rate, whereas the minimax optimal rate is $N^{-1/2}$ when $m = n$. We build on the observation that L_{CvM} converges to a normal distribution under both H_0 and H_1 to prove the following result.

Proposition C.7 (Non-optimality of the linear time test)**.** *Let $c_{\alpha,\text{linear}}$ be the α level critical value of the oracle test (see Section 4.2.2 of the main text) based on L_{CvM} in (C.55) and define the corresponding test function by*

$$\phi_{L_{\text{CvM}}} := \mathbb{1}(L_{\text{CvM}} > c_{\alpha,\text{linear}}).$$

Consider a sequence of alternatives such that

$$W_d(P_X, P_Y) \asymp N^{-\varepsilon} \quad \textit{where} \quad \varepsilon > 1/4.$$

Then for $0 < \alpha < 1/2$,

$$\lim_{m,n\to\infty} \mathbb{P}_1(\phi_{L_{\mathrm{CvM}}} = 1) \le 1/2.$$

Proof. Let σ_0^2 and σ_1^2 be the variance of

$$\widetilde{h}_{\mathrm{CvM}}(X_1, X_2; Y_1, Y_2) = \frac{1}{2}\{h_{\mathrm{CvM}}(X_1, X_2; Y_1, Y_2) + h_{\mathrm{CvM}}(X_2, X_1; Y_2, Y_1)\},$$

under the null and alternative, respectively. From the boundedness of h_{CvM}, we have $0 < \sigma_0^2, \sigma_1^2 < \infty$. Then by the central limit theorem, the null distribution approximates

$$\frac{\sqrt{M} L_{\mathrm{CvM}}}{\sigma_0} \xrightarrow{d} N(0,1) \quad \text{under } H_0,$$

which implies that $\sqrt{M}\sigma_0^{-1} c_{\alpha,\mathrm{linear}} \to -z_\alpha$ where z_α is the α quantile of the standard normal distribution and $z_\alpha < 0$ for $\alpha < 1/2$. Hence, the power function approximates

$$\begin{aligned}
& \lim_{N\to\infty} \mathbb{P}_1\left(L_{\mathrm{CvM}} > c_{\alpha,\mathrm{linear}}\right) \\
= & \lim_{N\to\infty} \mathbb{P}_1\left(\frac{\sqrt{M}(L_{\mathrm{CvM}} - W_d^2)}{\sigma_1} > \frac{\sqrt{M} c_{\alpha,\mathrm{linear}}}{\sigma_1} - \frac{\sqrt{M} W_d^2}{\sigma_1}\right) \\
= & \lim_{N\to\infty} \mathbb{P}_1\left(\frac{\sqrt{M}(L_{\mathrm{CvM}} - W_d^2)}{\sigma_1} > -\frac{\sigma_0}{\sigma_1} z_\alpha - \frac{\sqrt{M} W_d^2}{\sigma_1}\right) \\
\le & \lim_{N\to\infty} \mathbb{P}_1\left(\frac{\sqrt{M}(L_{\mathrm{CvM}} - W_d^2)}{\sigma_1} > -\frac{\sqrt{M} W_d^2}{\sigma_1}\right) \\
= & \frac{1}{2},
\end{aligned}$$

where the last equality uses

$$\frac{\sqrt{M}(L_{\mathrm{CvM}} - W_d^2)}{\sigma_1} \xrightarrow{d} N(0,1) \quad \text{under } H_1$$

and $\sqrt{M} W_d^2 \xrightarrow{p} 0$ by the assumption. This completes the proof. □

Continuing our discussion from Remark 4.7 in the main text, one can come up with different test statistics by modifying the CvM-distance. As explained there, H_0 holds if and only if $\mathsf{T}_{xy} = \mathsf{T}_{xx}$ and $\mathsf{T}_{xy} = \mathsf{T}_{yy}$ where $\mathsf{T}_{xy} = \mathbb{E}[\rho_{\mathrm{Angle}}(X_1, Y_1)]$, $\mathsf{T}_{xx} = \mathbb{E}[\rho_{\mathrm{Angle}}(X_1, X_2)]$ and $\mathsf{T}_{yy} = \mathbb{E}[\rho_{\mathrm{Angle}}(Y_1, Y_2)]$. Empirical estimates of

these quantities can be given by

$$\begin{aligned}
\widehat{\mathsf{T}}_{xy} &= \frac{1}{2\pi(m)_2 \cdot n} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j=1}^{n} \mathsf{Ang}(X_{i_1} - X_{i_2}, Y_j - X_{i_2}) \\
&\quad + \frac{1}{2\pi m \cdot (n)_2} \sum_{i=1}^{m} \sum_{j_1,j_2=1}^{n,\neq} \mathsf{Ang}(X_i - Y_{j_2}, Y_{i_1} - Y_{i_2}), \\
\widehat{\mathsf{T}}_{xx} &= \frac{1}{2\pi(m)_3} \sum_{i_1,i_2,i_3=1}^{m,\neq} \mathsf{Ang}(X_{i_1} - X_{i_3}, X_{i_2} - X_{i_3}) \\
&\quad + \frac{1}{2\pi(m)_2 \cdot n} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j=1}^{n} \mathsf{Ang}(X_{i_1} - Y_j, X_{i_2} - Y_j), \\
\widehat{\mathsf{T}}_{yy} &= \frac{1}{2\pi(n)_3} \sum_{j_1,j_2,j_3=1}^{n,\neq} \mathsf{Ang}(Y_{j_1} - Y_{j_3}, Y_{j_2} - Y_{j_3}) \\
&\quad + \frac{1}{2\pi m \cdot (n)_2} \sum_{i=1}^{m} \sum_{j_1,j_2=1}^{n,\neq} \mathsf{Ang}(Y_{j_1} - X_i, Y_{j_2} - X_i).
\end{aligned}$$

One possible test statistic that combines $\widehat{\mathsf{T}}_{xy}, \widehat{\mathsf{T}}_{xx}$ and $\widehat{\mathsf{T}}_{yy}$ is given by

$$V_{\mathrm{CvM}} = \big(\widehat{\mathsf{T}}_{xy} - \widehat{\mathsf{T}}_{xx}\big)^2 + \big(\widehat{\mathsf{T}}_{xy} - \widehat{\mathsf{T}}_{yy}\big)^2. \tag{C.56}$$

More generally, from Lemma C.1.1 and its proof, one can see that

$$W_d^2 = \vartheta_X(2\mathsf{T}_1 - \mathsf{T}_2 - \mathsf{T}_3) + \vartheta_Y(2\mathsf{T}_4 - \mathsf{T}_5 - \mathsf{T}_6),$$

where

$$\begin{aligned}
\mathsf{T}_1 &= \frac{1}{2\pi}\mathbb{E}[\mathsf{Ang}(X_1 - X_2, Y_1 - X_2)], \quad \mathsf{T}_2 = \frac{1}{2\pi}\mathbb{E}[\mathsf{Ang}(X_1 - X_3, X_2 - X_3)], \\
\mathsf{T}_3 &= \frac{1}{2\pi}\mathbb{E}[\mathsf{Ang}(Y_1 - X_1, Y_2 - X_1)], \quad \mathsf{T}_4 = \frac{1}{2\pi}\mathbb{E}[\mathsf{Ang}(X_1 - Y_2, Y_1 - Y_2)], \\
\mathsf{T}_5 &= \frac{1}{2\pi}\mathbb{E}[\mathsf{Ang}(X_1 - Y_1, X_2 - Y_1)], \quad \mathsf{T}_6 = \frac{1}{2\pi}\mathbb{E}[\mathsf{Ang}(Y_1 - Y_3, Y_2 - Y_3)].
\end{aligned}$$

As one of the reviewers pointed out, H_0 is true if and only if all the four equalities hold together: $\mathsf{T}_1 = \mathsf{T}_2$, $\mathsf{T}_1 = \mathsf{T}_3$, $\mathsf{T}_4 = \mathsf{T}_5$ and $\mathsf{T}_4 = \mathsf{T}_6$. Notice that $\mathsf{T}_1, \ldots, \mathsf{T}_6$ can be estimated in a straightforward manner as

$$\widehat{\mathsf{T}}_1 = \frac{1}{2\pi(m)_2 \cdot n} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j=1}^{n} \mathsf{Ang}(X_{i_1} - X_{i_2}, Y_j - X_{i_2}),$$

$$
\begin{aligned}
\widehat{\mathsf{T}}_2 &= \frac{1}{2\pi(m)_3} \sum_{i_1,i_2,i_3=1}^{m,\neq} \mathsf{Ang}(X_{i_1} - X_{i_3}, X_{i_2} - X_{i_3}), \\
\widehat{\mathsf{T}}_3 &= \frac{1}{2\pi m \cdot (n)_2} \sum_{i=1}^{m} \sum_{j_1,j_2=1}^{n,\neq} \mathsf{Ang}(Y_{j_1} - X_i, Y_{j_2} - X_i), \\
\widehat{\mathsf{T}}_4 &= \frac{1}{2\pi m \cdot (n)_2} \sum_{i=1}^{m} \sum_{j_1,j_2=1}^{n,\neq} \mathsf{Ang}(X_i - Y_{j_2}, Y_{i_1} - Y_{i_2}), \\
\widehat{\mathsf{T}}_5 &= \frac{1}{2\pi(m)_2 \cdot n} \sum_{i_1,i_2=1}^{m,\neq} \sum_{j=1}^{n} \mathsf{Ang}(X_{i_1} - Y_j, X_{i_2} - Y_j), \\
\widehat{\mathsf{T}}_6 &= \frac{1}{2\pi(n)_3} \sum_{j_1,j_2,j_3=1}^{n,\neq} \mathsf{Ang}(Y_{j_1} - Y_{j_3}, Y_{j_2} - Y_{j_3}).
\end{aligned}
$$

Based on these estimators, another test statistic, suggested by one of the reviewers, can be proposed as

$$
\begin{aligned}
V^*_{\mathrm{CvM}} &= \frac{m}{m+n}\Big\{\big(\widehat{\mathsf{T}}_1 - \widehat{\mathsf{T}}_2\big)^2 + \big(\widehat{\mathsf{T}}_1 - \widehat{\mathsf{T}}_3\big)^2\Big\} \\
&\quad + \frac{n}{m+n}\Big\{\big(\widehat{\mathsf{T}}_4 - \widehat{\mathsf{T}}_5\big)^2 + \big(\widehat{\mathsf{T}}_4 - \widehat{\mathsf{T}}_6\big)^2\Big\}.
\end{aligned} \tag{C.57}
$$

As demonstrated in Appendix C.6, the tests based on V_{CvM} and V^*_{CvM} tend to have higher power than that based on U_{CvM} against scale alternatives but lower power against location alternatives.

C.5.5 Power expression in HDLSS regime

Recall the five test statistics considered in Section 4.5 of the main text. This subsection provides an explicit expression for the limiting power function under the HDLSS regime. To this end, we need more restrictions on X and Y such as stationary ρ-mixing condition. Then we build on the asymptotic results established in Chakraborty and Chaudhuri (2017) combined with Theorem 4.11 to have the following corollary.

Corollary C.1.1 (Power of asymptotic tests)**.** *Consider the same assumptions made in Theorem 4.11. Assume that $X = \mu_X + V_X$ and $Y = \mu_Y + V_Y$ where $\mathbb{E}(V_X) = \mathbb{E}(V_Y) = 0$ and V_X and V_Y are mutually independent random vectors in $\mathbb{R}^d$. In addition, assume that the components of $V_X = (V_{X,1}, V_{X,2}, \ldots,)$ are strictly stationary and satisfy $\sum_{k=1}^{\infty} \rho_X(2^k) < \infty$ where $\rho_X(\cdot)$ is the ρ-mixing coefficient. The components of $V_Y = (V_{Y,1}, V_{Y,2}, \ldots,)$ are similarly defined with another mixing coefficient $\rho_Y(\cdot)$. Let $\{X_i\}_{i=1}^m$ be i.i.d. copies of X and $\{Y_i\}_{i=1}^n$ be i.i.d. copies of Y. Denote*

$$
\psi_{m,n} = \mathrm{tr}(\Sigma^2)\{2/m_{(2)} + 2/n_{(2)} + 4/(mn)\},
$$

and write the test functions by

$$\phi'_{\mathrm{CvM}} = \mathbb{1}(2\pi\sqrt{3}d\overline{\sigma}^2 U_{\mathrm{CvM}} > z_\alpha \psi_{m,n}^{1/2}),$$

$$\phi'_{\mathrm{Energy}} = \mathbb{1}(\sqrt{2d}\overline{\sigma} U_{\mathrm{Energy}} > z_\alpha \psi_{m,n}^{1/2}),$$

$$\phi'_{\mathrm{MMD}} = \mathbb{1}(\varsigma_d^2 e^{-d\overline{\sigma}_d^2/\varsigma_d^2} U_{\mathrm{MMD}} > z_\alpha \psi_{m,n}^{1/2}),$$

$$\phi'_{\mathrm{CQ}} = \mathbb{1}(U_{\mathrm{CQ}} > z_\alpha \psi_{m,n}^{1/2}) \quad \textit{and}$$

$$\phi'_{\mathrm{WMW}} = \mathbb{1}(d\overline{\sigma}^2 U_{\mathrm{WMW}} > z_\alpha \psi_{m,n}^{1/2}).$$

Then under the HDLSS setting,

$$\lim_{d\to\infty} \mathbb{E}[\phi'_{\mathrm{CvM}}] = \lim_{d\to\infty} \mathbb{E}[\phi'_{\mathrm{Energy}}] = \lim_{d\to\infty} \mathbb{E}[\phi'_{\mathrm{MMD}}] = \lim_{d\to\infty} \mathbb{E}[\phi'_{\mathrm{CQ}}] = \lim_{d\to\infty} \mathbb{E}[\phi'_{\mathrm{WMW}}],$$

which converges to

$$\Phi\Big(- z_\alpha + \psi_{m,n}^{-1/2} \|\mu_X - \mu_Y\|^2 \Big),$$

where z_α is the upper α quantile of the standard normal distribution.

Proof. Under the stated assumptions, Theorem 2.1 of Chakraborty and Chaudhuri (2017) is satisfied. Hence the results for the CQ and WMW tests follow. For the rest of the tests, we apply Slutsky's theorem combined with Theorem 4.11 to obtain the results. This completes the proof. □

C.5.6 Angular distance is a metric of negative-type

Recall that $\mathcal{M}_X$ and $\mathcal{M}_Y$ are the support of X and Y respectively and $\mathcal{M} = \mathcal{M}_X \cup \mathcal{M}_Y \subseteq \mathbb{R}^d$. The next lemma shows that ρ_{Angle} (see Definition 4.2) is a metric of negative type defined on $\mathcal{M}$.

Lemma C.1.31. *For $\forall z, z', z'' \in \mathcal{M}$ and $\rho_{Angle} : \mathcal{M} \times \mathcal{M} \mapsto [0, \infty)$, the following conditions are satisfied*

1. $\rho_{\mathrm{Angle}}(z, z') \geq 0$ *and* $\rho_{\mathrm{Angle}}(z, z') = 0$ *if and only if* $z = z'$.

2. $\rho_{\mathrm{Angle}}(z, z') = \rho_{\mathrm{Angle}}(z', z)$.

3. $\rho_{\mathrm{Angle}}(z, z') \leq \rho_{\mathrm{Angle}}(z, z'') + \rho_{\mathrm{Angle}}(z', z'')$.

In addition, for $\forall n \geq 2$, $z_1, \ldots, z_n \in \mathcal{M}$, *and* $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, *with* $\sum_{i=1}^n \alpha_i = 0$,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho_{\text{Angle}}(z_i, z_j) \leq 0.$$

Proof. For given $w \in \mathbb{R}^d$, it is seen that

$$\begin{aligned}
& \int_{\mathbb{S}^{d-1}} \left| \mathbb{1}(\beta^\top z \leq \beta^\top w) - \mathbb{1}(\beta^\top z' \leq \beta^\top w) \right| d\lambda(\beta) \qquad \text{(C.58)} \\
= & \int_{\mathbb{S}^{d-1}} \mathbb{1}(\beta^\top z \leq \beta^\top w < \beta^\top z') + \mathbb{1}(\beta^\top z' \leq \beta^\top w < \beta^\top z) d\lambda(\beta) \\
= & \frac{1}{2} - \frac{1}{2\pi} \arccos\left\{ \frac{(z-w)^\top (w-z')}{\|z-w\| \|w-z'\|} \right\} + \frac{1}{2} - \frac{1}{2\pi} \arccos\left\{ \frac{(z'-w)^\top (w-z)}{\|z'-w\| \|w-z\|} \right\} \\
= & 1 - \frac{1}{\pi} \arccos\left\{ \frac{(z-w)^\top (w-z')}{\|z-w\| \|w-z'\|} \right\} \\
= & \frac{1}{\pi} \left(\pi - \arccos\left\{ \frac{(z-w)^\top (w-z')}{\|z-w\| \|w-z'\|} \right\} \right) \\
\overset{(i)}{=} & \frac{1}{\pi} \arccos\left\{ \frac{(z-w)^\top (z'-w)}{\|z-w\| \|z'-w\|} \right\} := \rho_{\text{Angle}}(z, z'; w),
\end{aligned}$$

where (i) is due to $\arccos(x) + \arccos(-x) = \pi$. Then $\rho_{\text{Angle}}(z, z')$ is the expected value of $\rho_{\text{Angle}}(z, z'; Z^*)$ over $Z^* \sim (1/2)P_X + (1/2)P_Y$, i.e.

$$\begin{aligned}
\rho_{\text{Angle}}(z, z') &= \mathbb{E}\left[\rho_{\text{Angle}}(z, z'; Z^*) \right] \\
&= \frac{1}{\pi} \mathbb{E}\left[\arccos\left\{ \frac{(z-Z^*)^\top (z'-Z^*)}{\|z-Z^*\| \|z'-Z^*\|} \right\} \right].
\end{aligned}$$

Now, if $z = z'$, it is trivial to see $\rho_{\text{Angle}}(z, z') = 0$. In addition, if $\rho_{\text{Angle}}(z, z') = 0$, then we have $z = z'$. In order to show the second direction, note that $\arccos(x)$ is positive and monotone decreasing over $x \in [-1, 1]$ and so $\rho_{\text{Angle}}(z, z') = 0$ implies that

$$\frac{(z-Z^*)^\top (z'-Z^*)}{\|z-Z^*\| \|z'-Z^*\|} = 1,$$

almost surely with respect to $(1/2)P_X + (1/2)P_Y$. By Cauchy-Schwarz inequality, the inner product becomes one if and only if $(z - Z^*)$ or $(z' - Z^*)$ is a multiple of the other. This is only possible when $z - Z^* = z' - Z^*$ almost surely, which implies $z = z'$. The symmetry property follows easily by the definition of ρ_{Angle}. In

addition, from triangle inequality, we have

$$\begin{aligned}
&\int_{\mathbb{S}^{d-1}} \Big|\mathbb{1}(\beta^\top z \le \beta^\top w) - \mathbb{1}(\beta^\top z' \le \beta^\top w)\Big| d\lambda(\beta) \\
\le\ &\int_{\mathbb{S}^{d-1}} \Big|\mathbb{1}(\beta^\top z \le \beta^\top w) - \mathbb{1}(\beta^\top z'' \le \beta^\top w)\Big| d\lambda(\beta) + \int_{\mathbb{S}^{d-1}} \Big|\mathbb{1}(\beta^\top z'' \le \beta^\top w) - \mathbb{1}(\beta^\top z' \le \beta^\top w)\Big| d\lambda(\beta),
\end{aligned}$$

and therefore by the equality in (C.58), we can establish

$$\rho_{\text{Angle}}(z, z'; w) \ \le\ \rho_{\text{Angle}}(z, z''; w) + \rho_{\text{Angle}}(z', z''; w).$$

Now, by taking the expectation over Z^*, we conclude that

$$\rho_{\text{Angle}}(z, z') \ \le\ \rho_{\text{Angle}}(z, z'') + \rho_{\text{Angle}}(z', z'').$$

Next, we will show that for $\forall n \ge 2$, $z_1, \ldots, z_n \in S$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^n \alpha_i = 0$,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho_{\text{Angle}}(z_i, z_j) \le 0.$$

The result follows from Section 6 of Bogomolny et al. (2007) who showed that for each fixed z^*,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho_{\text{Angle}}(z_i, z_j; z^*) \le 0, \tag{C.59}$$

for any $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^n \alpha_i = 0$. Therefore, by taking the expected value over z^* in (C.59), we conclude that ρ_{Angle} is of negative-type. □

C.5.7 Details on Remark 4.6

Regarding Remark 4.6, note that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \rho_{\text{Angle}}(z, z'; t) dt \\
=\ &\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} I(\beta^\top z \le \beta^\top t < \beta^\top z') + \mathbb{1}(\beta^\top z' \le \beta^\top t < \beta^\top z) d\beta^\top t d\lambda(\beta) \\
\overset{(i)}{=}\ &\int_{\mathbb{S}^{d-1}} |\beta^\top (z - z')| d\lambda(\beta) \\
\overset{(ii)}{=}\ &\gamma_d \|z - z'\|,
\end{aligned}$$

where (i) and (ii) are due to Lemma 2.1 and Lemma 2.3 of Baringhaus and Franz (2004) and

$$\gamma_d = \frac{\sqrt{\pi}(d-1)\Gamma\left((d-2)/2\right)}{2\Gamma(d/2)}.$$

Therefore, the generalized angular distance with Lebesgue measure corresponds to the Euclidean distance.

C.5.8 Further applications of projection-averaging

Given $(X_i, Y_i) \overset{i.i.d.}{\sim} P_{XY}$, recall that Kendall's tau is an estimate of $4\mathbb{P}(X_1 < X_2, Y_1 < Y_2) - 1$. Another common measure of association is Spearman's rho defined by

$$\rho^{\text{sp}} = 12\mathbb{P}(X_1 < X_2, Y_1 < Y_3) - 3.$$

Now for $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, its multivariate extension via projection-averaging can be given by

$$\rho^{\text{sp}}_{p,q} = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{q-1}} \left[12\mathbb{P}(\alpha^\top X_1 < \alpha^\top X_2, \beta^\top Y_1 < \beta^\top Y_3) - 3\right]^2 d\lambda(\alpha) d\lambda(\beta).$$

The next proposition gives a closed-form expression for $\rho_{p,q}$ via Lemma 4.0.2.

Proposition C.8 (Spearman's rho)**.** *For i.i.d. pairs of random vectors $(X_1, Y_1), \ldots, (X_6, Y_6)$ from a joint distribution P_{XY} where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, the multivariate extension of Spearman's rho via projection-averaging is given by*

$$\rho^{\text{sp}}_{p,q} = 144\mathbb{E}\left[\left(\frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}(X_1 - X_2, X_4 - X_5)\right) \times \left(\frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}(Y_1 - Y_3, Y_5 - Y_6)\right)\right] - 9.$$

Proof. Given $\alpha \in \mathbb{S}^{p-1}, \beta \in \mathbb{S}^{q-1}$, expand the square term of $\rho^{\text{sp}}_{p,q}$ to have

$$\begin{aligned}
&\left[12\mathbb{P}(\alpha^\top X_1 < \alpha^\top X_2, \beta^\top Y_1 < \beta^\top Y_3) - 3\right]^2 \\
&= 144\mathbb{E}\left[\mathbb{1}\{\alpha^\top(X_1 - X_2) < 0\}\mathbb{1}\{\beta^\top(Y_1 - Y_3) < 0\} \times \mathbb{1}\{\alpha^\top(X_4 - X_5) < 0\}\mathbb{1}\{\beta^\top(Y_4 - Y_6) < 0\}\right] \\
&\quad -72\mathbb{E}\left[\mathbb{1}\{\alpha^\top(X_1 - X_2) < 0\}\mathbb{1}\{\beta^\top(Y_1 - Y_3) < 0\}\right] + 9.
\end{aligned}$$

Then applying Lemma 4.0.2 with Fubini's theorem yields the expression. □

For a multivariate case, Zhu et al. (2017) extended Hoeffding's coefficient (Hoeffding, 1948) via projection-averaging. Specifically, they defined the projection correlation between $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ as

$$\int_{\mathbb{S}^{p-1}} \int_{\mathbb{S}^{q-1}} \int_{\mathbb{R}^2} \left[F_{\alpha^\top X, \beta^\top Y}(u, v) - F_{\alpha^\top X}(u) F_{\beta^\top Y}(v)\right]^2 d\omega_1(u, v, \alpha, \beta), \tag{C.60}$$

where $d\omega_1(u,v,\alpha,\beta) = dF_{\alpha^\top X,\beta^\top Y}(u,v)d\lambda(\alpha)d\lambda(\beta)$. Although the projection correlation is more broadly sensitive than Kendall's tau or Pearson's correlation in detecting dependence among random variables, it can still be zero even when X and Y are dependent. A counterexample for the univariate case can be found in Hoeffding (1948).

On the other hand, the coefficient introduced by Blum et al. (1961) overcomes this issue by replacing $dF_{X,Y}$ with $dF_X dF_Y$. The univariate Blum–Kiefer–Rosenblatt (BKR) coefficient (Blum et al., 1961) is defined by

$$\int_{\mathbb{R}^2} \left[F_{XY}(u,v) - F_X(u)F_Y(v)\right]^2 dF_X(u)dF_Y(v).$$

Leveraging Lemma 4.0.2, the next proposition generalizes the univariate BKR coefficient to a multivariate space.

Proposition C.9 (Blum–Kiefer–Rosenblatt (BKR) coefficient)**.** *Let us consider weight function* $d\omega_2(u,v,\alpha,\beta) = dF_{\alpha^\top X}(u)dF_{\beta^\top Y}(v)d\lambda(\alpha)d\lambda(\beta)$. *For i.i.d. random vectors* $(X_1,Y_1),\ldots,(X_6,Y_6)$ *from a joint distribution* P_{XY} *where* $X \in \mathbb{R}^p$ *and* $Y \in \mathbb{R}^q$*, the univariate BKR coefficient can be extended to a multivariate case by*

$$\begin{aligned}
&\int_{\mathbb{S}^{p-1}}\int_{\mathbb{S}^{q-1}}\int_{\mathbb{R}^2} \left[F_{\alpha^\top X,\beta^\top Y}(u,v) - F_{\alpha^\top X}(u)F_{\beta^\top Y}(v)\right]^2 d\omega_2(u,v,\alpha,\beta)\\
=\ &\mathbb{E}\left[\left(\frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}\left(X_1 - X_3, X_2 - X_3\right)\right)\cdot\left(\frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}\left(Y_1 - Y_4, Y_2 - Y_4\right)\right)\right]\\
+\ &\mathbb{E}\left[\left(\frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}\left(X_1 - X_5, X_2 - X_5\right)\right)\cdot\left(\frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}\left(Y_3 - Y_6, Y_4 - Y_6\right)\right)\right]\\
-2\ &\mathbb{E}\left[\left(\frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}\left(X_1 - X_4, X_2 - X_4\right)\right)\cdot\left(\frac{1}{2} - \frac{1}{2\pi}\mathsf{Ang}\left(Y_1 - Y_5, Y_3 - Y_5\right)\right)\right].
\end{aligned}$$

Proof. Given $\alpha \in \mathbb{S}^{p-1}$ and $\beta \in \mathbb{S}^{q-1}$,

$$\begin{aligned}
&\int_{\mathbb{R}^2} \left[F_{\alpha^\top X,\beta^\top Y}(u,v) - F_{\alpha^\top X}(u)F_{\beta^\top Y}(v)\right]^2 dF_{\alpha^\top X}(u)dF_{\beta^\top Y}(v)\\
=\ &\mathbb{E}\Big[\mathbb{1}(\alpha^\top(X_1 - X_3) \le 0, \alpha^\top(X_2 - X_3) \le 0) \times \mathbb{1}(\beta^\top(Y_1 - Y_4) \le 0, \beta^\top(Y_2 - Y_4) \le 0)\Big]\\
+\ &\mathbb{E}\Big[\mathbb{1}(\alpha^\top(X_1 - X_5) \le 0, \alpha^\top(X_2 - X_5) \le 0) \times \mathbb{1}(\beta^\top(Y_3 - Y_6) \le 0, \beta^\top(Y_4 - Y_6) \le 0)\Big]\\
-2\mathbb{E}\Big[&\mathbb{1}(\alpha^\top(X_1 - X_4) \le 0, \alpha^\top(X_2 - X_4) \le 0) \times \mathbb{1}(\beta^\top(Y_1 - Y_5) \le 0, \beta^\top(Y_3 - Y_5) \le 0)\Big].
\end{aligned}$$

Then apply Lemma 4.0.2 to obtain the expression. □

We end this subsection with one example that motivates Lemma C.1.30. Notice that most of the univariate test statistics extended via projection-averaging are based on a squared difference between

distribution functions. In such a case, Lemma 4.0.2 is enough to obtain their multivariate extensions. Suppose that one is interested in the fourth power of difference between two distribution functions:

$$\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left[F_{\beta^\top X}(t) - F_{\beta^\top Y}(t)\right]^4 dH_\beta(t) d\lambda(\beta).$$

In this case, however, it is necessary to use the explicit formula for the integration involving four indicator functions. For example, by expanding the fourth power, we may have different terms including

$$\begin{aligned} &\int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \left(F_{\beta^\top X}(t)\right)^4 dF_{\beta^\top X}(t) d\lambda(\beta) \\ =\ &\mathbb{E}\big[\mathbb{1}(\beta^\top(X_1 - X_5) \le 0)\mathbb{1}(\beta^\top(X_2 - X_5) \le 0) \times \mathbb{1}(\beta^\top(X_3 - X_5) \le 0)\mathbb{1}(\beta^\top(X_4 - X_5) \le 0)\big]. \end{aligned}$$

This expectation requires the integration over $\mathbb{S}^{d-1}$, but one can apply Lemma C.1.30 to reduce it to one-dimensional integration, which is much more tractable in practice.

C.6 Additional Simulations

This section provides additional simulation results to further evaluate the performance of the considered tests under different scenarios.

C.6.1 High-dimensional power under strong dependence

Let us first consider the setting where the component variables are strongly dependent. Specifically, we assume that X has a multivariate t-distribution with the location parameter $\mu_X = (0, \ldots, 0)^\top$, the degrees of freedom υ and the $d \times d$ shape matrix S where $[S]_{ij} = 1$ if $i = j$ and $[S]_{ij} = 0.9$ otherwise. Note that when $\upsilon > 2$, the covariance matrix of X is given by $\frac{\upsilon}{\upsilon - 2} S$. Similarly, we assume that Y has a multivariate t-distribution with the location parameter $\mu_X = (0.2, \ldots, 0.2)^\top$, the degrees of freedom υ and the shape matrix S. Under the given setting, we generated $m = n = 20$ random samples from each distribution with $d = 200$ and carried out the permutation tests as in Section 8.9. We increased the degrees of freedom from $\upsilon = 1$ to $\upsilon = \infty$ to vary the moment conditions. As shown in Table C.1, the WMW test performs the best when $\upsilon \le 7$ closely followed by the CvM test. When υ is large (e.g. $\upsilon \ge 20$) meaning that X and Y have relatively light-tailed distributions, the power of the five tests (CvM, Energy, MMD, CQ, WMW) are very similar as observed in Section 8.9. These empirical results provide evidence that the findings in Section 4.5 may hold under even more general settings where the component variables are strongly dependent.

Table C.1: Empirical power of the considered tests at $\alpha = 0.05$ against the location models when the component variables are strongly dependent.

$m = 20, n = 20$	$\upsilon = 1$	$\upsilon = 3$	$\upsilon = 5$	$\upsilon = 7$	$\upsilon = 9$	$\upsilon = 11$	$\upsilon = 20$	$\upsilon = \infty$
CvM	0.118	**0.653**	**0.823**	**0.880**	**0.907**	**0.918**	**0.943**	**0.943**
Energy	0.053	0.332	**0.642**	**0.808**	**0.865**	**0.887**	**0.937**	**0.945**
MMD	0.075	0.162	0.363	0.595	0.755	0.810	**0.923**	**0.945**
CQ	0.063	0.470	**0.692**	**0.815**	**0.842**	**0.892**	**0.920**	**0.943**
WMW	**0.340**	**0.767**	**0.865**	**0.892**	**0.892**	**0.930**	**0.942**	**0.943**
NN	**0.293**	**0.490**	0.528	0.532	0.528	0.533	0.577	0.583
FR	**0.225**	0.322	0.305	0.313	0.307	0.293	0.283	0.378
MBG	0.047	0.062	0.053	0.043	0.048	0.052	0.050	0.100
Ball	0.063	0.050	0.057	0.053	0.070	0.070	0.075	0.620
CM	0.052	0.067	0.057	0.057	0.065	0.075	0.093	0.125
BG	0.040	0.045	0.047	0.040	0.065	0.048	0.058	0.185
Run	0.112	0.112	0.155	0.152	0.167	0.187	0.198	0.325

C.6.2 Low-dimensional Gaussian alternatives

Next we compare low-dimensional Gaussian distributions with different location or scale parameters. Suppose that X has a multivariate Gaussian distribution with zero mean $\mu_X = (0, \ldots, 0)^\top$ and identity covariance matrix $\Sigma_X = I_d$. Similarly Y has a multivariate Gaussian distribution with mean $\mu_Y = (\mu, \ldots, \mu)^\top$ and a diagonal covariance matrix $\Sigma_Y = \sigma \times I_d$. For the location alternatives, we fix the scale parameter $\sigma = 1$ and change the value of μ, whereas for the scale alternatives, we fix the location parameter $\mu = 0$ and change the value of σ. Throughout the simulation study, we set the dimension to be either $d = 2$ or $d = 8$ while the sample sizes are $m = n = 40$. In this simulation study, we consider the CvM test, the NN test and three other tests based on the modified test statistics L_{CvM}, V_{CvM} and V^*_{CvM} in (C.55), (C.56) and (C.57), respectively. We also add Hotelling's test (e.g. page 188 of Anderson, 2003) and the LRT test (e.g. page 412 of Anderson, 2003) as a reference point for the location alternative and the scale alternative, respectively. All tests were carried out by the permutation procedure as in Section 8.9. The simulation results are given in Figure C.1 and Figure C.2.

Let us first look at the power of the tests under mean differences in Figure C.1. What stands out in this figure is that the CvM test has approximately the same power as Hotelling's test, which has a certain optimality property under the considered location alternatives (e.g. Chapter 5 of Anderson, 1962). However we should emphasize that, unlike Hotelling's test, the CvM test can have power against much broader alternatives. The modified CvM tests based on V_{CvM} and V^*_{CvM} show a good power performance against the same location alternatives and they are slightly more powerful than the NN test but less powerful than

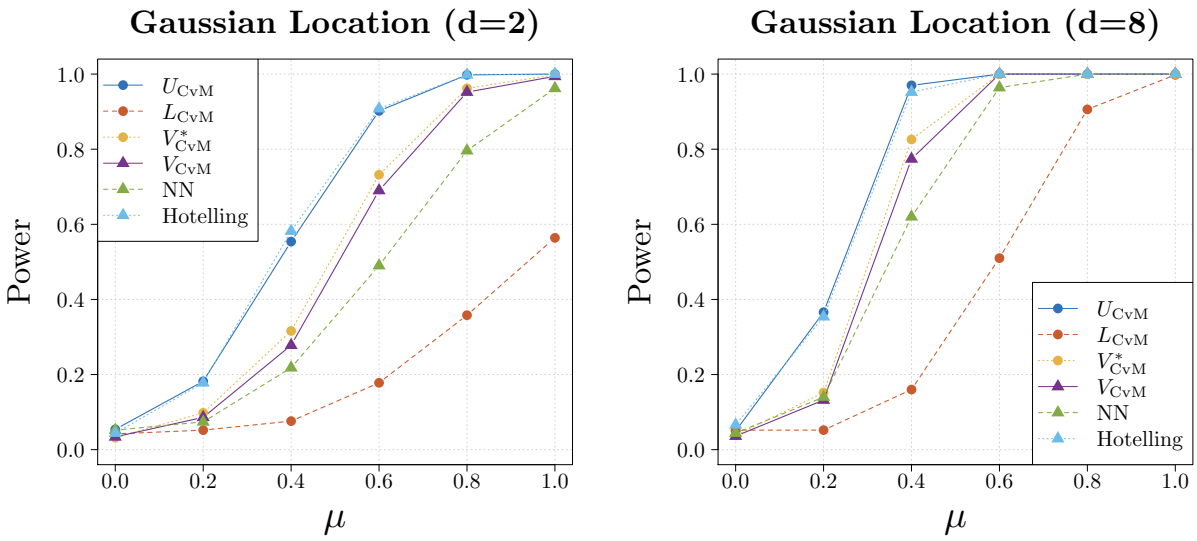

Figure C.1: Empirical power of the considered tests at $\alpha = 0.05$ under Gaussian location alternatives.

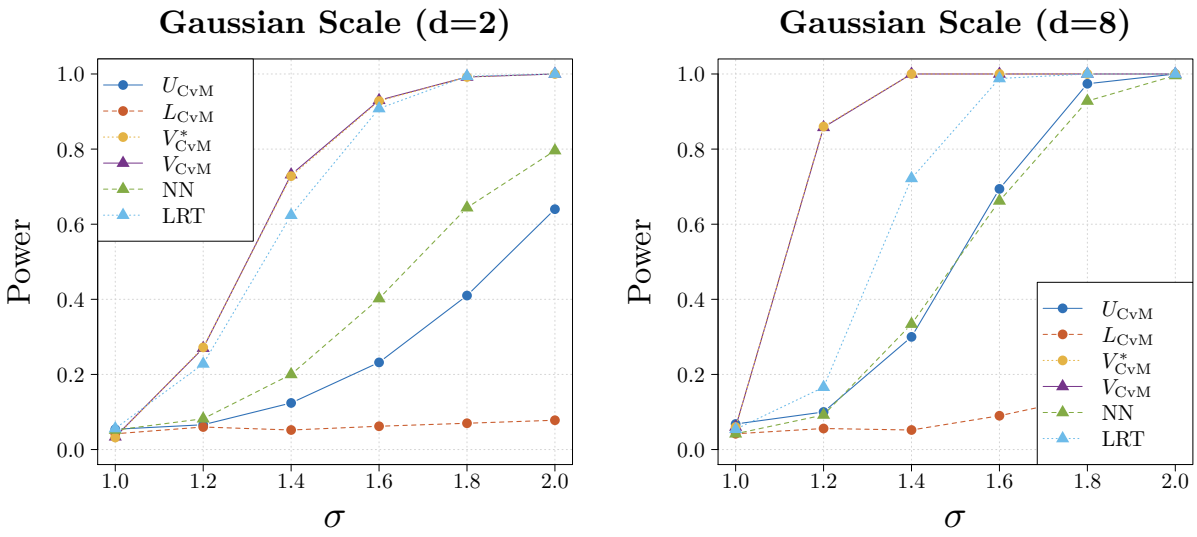

Figure C.2: Empirical power of the considered tests at $\alpha = 0.05$ under Gaussian scale alternatives.

the original CvM test. Interestingly these modified tests observe a huge power enhancement against scale differences in Figure C.2, outperforming the other nonparametric tests. In fact they perform comparable to or even better than the LRT test as the dimension increases. The original CvM test tends to have lower or comparable power to the NN test against these light-tailed scale alternatives. The empirical results also confirm Proposition C.7, which shows that the test based on the linear-type statistic suffers from low power while having a linear time complexity.

Appendix D

Appendix for Chapter 5

D.1 Proofs

In this section, we collect the proofs of the theorems in the main text. Throughout this section, we use $C_1, C_2, \ldots$ to denote some constants that may change from line to line.

D.1.1 Proof of Theorem 5.1

The following proof is built upon the proof of Theorem 4.1 of Drton et al. (2018) and extends theirs to two-sample V-statistics and unbounded eigenfunctions. We start with another representation of $\widehat{\mathcal{V}}_{12}^2$ in terms of $\{\lambda_v\}_{v=1}^{\infty}$ and $\{\varphi_v(\cdot)\}_{v=1}^{\infty}$. Since $h(z_1, z_2)$ is symmetric in its arguments, $\widehat{\mathcal{V}}_{12}^2$ can also be represented in terms of the centered kernel as

$$\widehat{\mathcal{V}}_{12}^2 = \frac{1}{n_1^2}\sum_{i_1,i_2=1}^{n_1}\overline{h}(X_{i_1,1}, X_{i_2,1}) + \frac{1}{n_2^2}\sum_{i_1,i_2=1}^{n_2}\overline{h}(X_{i_1,2}, X_{i_2,2}) - \frac{2}{n_1 n_2}\sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2}\overline{h}(X_{i_1,1}, X_{i_2,2}).$$

Furthermore, based on the decomposition given in (5.2), $\widehat{\mathcal{V}}_{12}^2$ can be written as

$$\widehat{\mathcal{V}}_{12}^2 \;=\; \sum_{v=1}^{\infty}\lambda_v\left\{\frac{1}{n_1}\sum_{i_1=1}^{n_1}\varphi_v(X_{i_1,1}) - \frac{1}{n_2}\sum_{i_2=1}^{n_2}\varphi_v(X_{i_2,2})\right\}^2.$$

In what follows, we consider two different cases: 1) x is bounded and 2) x tends to infinity and prove Theorem 5.1 under each scenario.

Case 1: x is bounded

First write the corresponding degenerate two-sample U-statistic by

$$\widehat{\mathcal{U}}_{12} = \frac{1}{n_1(n_1-1)} \sum_{1\le i_1\neq i_2\le n_1} \overline{h}(X_{i_1,1}, X_{i_2,1}) + \frac{1}{n_2(n_2-1)} \sum_{1\le i_1\neq i_2\le n_2} \overline{h}(X_{i_1,2}, X_{i_2,2}) - \frac{2}{n_1 n_2} \sum_{i_1=1}^{n_1}\sum_{i_2=1}^{n_2} \overline{h}(X_{i_1,1}, X_{i_2,2}).$$

Then using the result on Chapter 3 of Bhat (1995),

$$\frac{n_1 n_2}{N}\widehat{\mathcal{U}}_{12} \xrightarrow{d} \sum_{v=1}^{\infty} \lambda_v(\xi_v^2 - 1).$$

Now the difference between the V-statistic and U-statistic is

$$\widehat{\mathcal{V}}_{12}^2 - \widehat{\mathcal{U}}_{12} = \frac{1}{n_1^2}\sum_{i_1=1}^{n_1} \overline{h}(X_{i_1,1}, X_{i_1,1}) + \frac{1}{n_2^2}\sum_{i_2=1}^{n_2} \overline{h}(X_{i_2,2}, X_{i_2,2}) - \frac{1}{n_1^2(n_1-1)} \sum_{1\le i_1\neq i_2\le n_1} \overline{h}(X_{i_1,1}, X_{i_2,1}) - \frac{1}{n_2^2(n_2-1)} \sum_{1\le i_1\neq i_2\le n_2} \overline{h}(X_{i_1,2}, X_{i_2,2}).$$

Under the assumption that $\mathbb{E}[|\overline{h}(X_1, X_1)|] < \infty$, we apply the strong law of large numbers for U-statistics (e.g. Theorem A of Section 5.4 in Serfling, 1980) to have

$$\frac{n_1 n_2}{N}\left(\widehat{\mathcal{V}}_{12}^2 - \widehat{\mathcal{U}}_{12}\right) \xrightarrow{a.s} \mathbb{E}[\overline{h}(X_1, X_1)] = \sum_{v=1}^{\infty} \lambda_v.$$

Hence we establish that

$$\frac{n_1 n_2}{N}\widehat{\mathcal{V}}_{12}^2 \xrightarrow{d} \sum_{v=1}^{\infty} \lambda_v \xi_v^2,$$

which leads to (5.4) for any bounded x.

Case 2: x tends to infinity

Next we focus on the case where x tends to infinity at a certain rate. To start, for a sufficiently large positive integer T to be specified later, let us define the truncated statistic

$$\widehat{\mathcal{V}}_T^2 = \sum_{v=1}^{T} \lambda_v \left\{ \frac{1}{\sqrt{n_1}} \sum_{i=1}^{n_1} \varphi_v(X_{i,1}) - \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} \varphi_v(X_{i,2}) \right\}^2.$$

Based on Slutsky's argument,

$$\begin{aligned}\mathbb{P}\left(\frac{n_1n_2}{N}\widehat{\mathcal{V}}_{12}^2 \geq x\right) &\leq \mathbb{P}\left(\frac{n_1n_2}{N}\widehat{\mathcal{V}}_{T}^2 \geq x-\epsilon_1\right)+\mathbb{P}\left\{\left|\frac{n_1n_2}{N}\left(\widehat{\mathcal{V}}_{12}^2-\widehat{\mathcal{V}}_{T}^2\right)\right| \geq \epsilon_1\right\} \\ &:= (I)+(II) \quad \text{(say)}.\end{aligned}$$

Here and hereafter $\epsilon_1, \epsilon_2, \epsilon_3$ are some positive constants that will be specified later. Let us rewrite

$$\sqrt{\frac{\lambda_v}{n_1}}\sum_{i_1=1}^{n_1}\varphi_v(X_{i_1,1})-\sqrt{\frac{\lambda_v}{n_2}}\sum_{i_2=1}^{n_2}\varphi_v(X_{i_2,2})=\sum_{i=1}^{N}\sqrt{\lambda_v}w_i\varphi_v(Z_i)$$

where

$$\begin{aligned}(w_1,\ldots,w_N) &= (n_1^{-1/2},\ldots,n_1^{-1/2},-n_2^{-1/2},\ldots,-n_2^{-1/2}), \\ (Z_1,\ldots,Z_N) &= (X_{1,1},\ldots,X_{n_1,1},X_{1,2},\ldots,X_{n_2,2}).\end{aligned}$$

Further let $\varphi^{\lambda}_{1,\ldots,T}(Z_i) = (\sqrt{\lambda_1}w_i\varphi_1(Z_i),\ldots,\sqrt{\lambda_T}w_i\varphi_T(Z_i))^\top$. For each $i=1,\ldots,N$, we verify the multivariate Bernstein condition used in Zaitsev (1987). Specifically, for any $u,v\in\mathbb{R}^T$ and $m=3,4,\ldots$, we have that

$$\begin{aligned}&\left|\mathbb{E}\left[\{\varphi^{\lambda}_{1\ldots T}(Z_i)^\top u\}^2\{\varphi^{\lambda}_{1\ldots T}(Z_i)^\top v\}^{m-2}\right]\right| \\ \overset{(i)}{\leq}& \left(\sqrt{\frac{\lambda_1}{n_1}}+\sqrt{\frac{\lambda_1}{n_2}}\right)^m\left|\mathbb{E}\left[\{\varphi_{1\ldots T}(Z_i)^\top u\}^2\{\varphi_{1\ldots T}(Z_i)^\top v\}^{m-2}\right]\right| \\ \overset{(ii)}{\leq}& \left(\sqrt{\frac{\lambda_1}{n_1}}+\sqrt{\frac{\lambda_1}{n_2}}\right)^m\gamma^m m^{m/2}\|u\|_2^2\|v\|_2^{m-2} \\ \overset{(iii)}{\leq}& \left(\sqrt{\frac{\lambda_1}{n_1}}+\sqrt{\frac{\lambda_1}{n_2}}\right)^m\gamma^m m!\|v\|_2^{m-2}\mathbb{E}\left[\{\varphi_{1\ldots T}(Z_i)^\top u\}^2\right]\end{aligned}$$

where

- (i) follows since $\lambda_1\leq\lambda_2\leq\ldots$ and $\max\{1/\sqrt{n_1},1/\sqrt{n_2}\}\leq 1/\sqrt{n_1}+1/\sqrt{n_2}$.
- (ii) uses the condition **(A2)**.
- (iii) uses $m!\geq m^{m/2}$ for all $m\geq 3$ and $\mathbb{E}\left[\{\varphi_{1\ldots T}(Z_i)^\top u\}^2\right]=\|u\|_2^2$.

Thus together with the assumption that $C_1^{-1}\leq n_1/n_2\leq C_1$, the multivariate Bernstein condition in Zaitsev (1987) is fulfilled with his notation $\tau = C_2N^{-1/2}$ for sufficiently large C_2. Consequently, we can apply

Theorem 1.1 of Zaitsev (1987) to show that

$$\mathbb{P}\left(\frac{n_1 n_2}{N}\widehat{\mathcal{V}}_T^2 \geq x - \epsilon_1\right) \leq \mathbb{P}\left[\sum_{v=1}^{T} \lambda_v \xi_v^2 \geq \{(x-\epsilon_1)^{1/2} - \epsilon_2\}^2\right] + C_3 T^{5/2} \exp\left(-\frac{\sqrt{N}\epsilon_2}{C_4 T^{5/2}}\right).$$

By applying Slutsky's argument again, the first term is bounded by

$$\mathbb{P}\left[\sum_{v=1}^{T} \lambda_v \xi_v^2 \geq \{(x-\epsilon_1)^{1/2} - \epsilon_2\}^2\right] \leq \mathbb{P}\left[\sum_{v=1}^{\infty} \lambda_v \xi_v^2 \geq \{(x-\epsilon_1)^{1/2} - \epsilon_2\}^2 - \epsilon_3\right] + \mathbb{P}\left(\left|\sum_{v=T+1}^{\infty} \lambda_v \xi_v^2\right| \geq \epsilon_3\right).$$

For a random variable X, let us denote the sub-Gaussian norm and sub-exponential norm by $\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\}$ and $\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\}$, respectively. By the property of the norm, Example 2.5.8 of Vershynin (2018) and Lemma 2.7.6 of Vershynin (2018), we observe that

$$\left\|\sum_{v=T+1}^{\infty} \lambda_v \xi_v^2\right\|_{\psi_1} \leq \sum_{v=T+1}^{\infty} \lambda_v \left\|\xi_v^2\right\|_{\psi_1} = \sum_{v=T+1}^{\infty} \lambda_v \left\|\xi_v\right\|_{\psi_2}^2 \leq C_5 \sum_{v=T+1}^{\infty} \lambda_v.$$

Then by Proposition 2.7.1 of Vershynin (2018),

$$\mathbb{P}\left(\left|\sum_{v=T+1}^{\infty} \lambda_v \xi_v^2\right| \geq \epsilon_3\right) \leq 2\exp\left(-\frac{\epsilon_3}{C_5 \sum_{v=T+1}^{\infty} \lambda_v}\right).$$

Thus

$$(I) \leq \mathbb{P}\left[\sum_{v=1}^{\infty} \lambda_v \xi_v^2 \geq \{(x-\epsilon_1)^{1/2} - \epsilon_2\}^2 - \epsilon_3\right] + C_3 T^{5/2} \exp\left(-\frac{\sqrt{N}\epsilon_2}{C_4 T^{5/2}}\right) + 2\exp\left(-\frac{\epsilon_3}{C_5 \sum_{v=T+1}^{\infty} \lambda_v}\right).$$

Next we focus on the term (II). Note that the multivariate moment condition in (5.3) implies the univariate sub-Gaussian condition for $\varphi_v(X_1)$ and $v = 1, 2, \ldots$. That is, there exists a constant $C_6 > 0$ independent of v such that

$$\mathbb{E}[\{\varphi_v(X_1)\}^m] \leq C_6 m^{m/2} \mathbb{E}[\{\varphi_v(X_1)\}^2] = C_6 m^{m/2} \quad \text{for all } m \geq 1.$$

Thus, followed by Proposition 2.7.1 of Vershynin (2018), $\varphi_v(X_1)$ has a finite sub-Gaussian norm and furthermore $\sup_{v\geq 1}\|\varphi_v(X_1)\|_{\psi_2} := C_7 < \infty$. Then

$$\begin{aligned}
\left\|\frac{n_1n_2}{N}\left(\widehat{\mathcal{V}}^2-\widehat{\mathcal{V}}_T^2\right)\right\|_{\psi_1} &\overset{(i)}{\leq} \frac{n_1n_2}{N}\sum_{v=T+1}^{\infty}\lambda_v\left\|\left[\frac{1}{n_1}\sum_{i_1=1}^{n_1}\varphi_v(X_{i_1,1})-\frac{1}{n_2}\sum_{i_2=1}^{n_2}\varphi_v(X_{i_2,2})\right]^2\right\|_{\psi_1} \\
&\overset{(ii)}{=} \frac{n_1n_2}{N}\sum_{v=T+1}^{\infty}\lambda_v\left\|\frac{1}{n_1}\sum_{i_1=1}^{n_1}\varphi_v(X_{i_1,1})-\frac{1}{n_2}\sum_{i_2=1}^{n_2}\varphi_v(X_{i_2,2})\right\|_{\psi_2}^2 \\
&\overset{(iii)}{\leq} C_8\frac{n_1n_2}{N}\sum_{v=T+1}^{\infty}\lambda_v\left[\frac{1}{n_1}\sum_{i_1=1}^{n_1}\left\|\varphi_v(X_{i_1,1})\right\|_{\psi_2}^2+\frac{1}{n_2}\sum_{i_2=1}^{n_2}\left\|\varphi_v(X_{i_2,2})\right\|_{\psi_2}^2\right] \\
&\overset{(iv)}{\leq} C_9\sum_{v=T+1}^{\infty}\lambda_v
\end{aligned}$$

where

- (i) uses the triangle inequality.
- (ii) uses Lemma 2.7.6 of Vershynin (2018).
- (iii) holds by Proposition 2.6.1 of Vershynin (2018).
- (iv) follows since $\sup_{v\geq 1}\|\varphi_v(X_1)\|_{\psi_2} < \infty$.

Based on the above result, we apply Markov's inequality to bound

$$(II) \;\leq\; \exp\left(-\frac{\epsilon_1}{C_9\sum_{v=T+1}^{\infty}\lambda_v}\right).$$

To summarize, we have obtain that

$$\begin{aligned}
\frac{\mathbb{P}(n_1n_2\widehat{\mathcal{V}}^2/N\geq x)}{\mathbb{P}(\sum_{v=1}^{\infty}\lambda_v\xi_v^2\geq x)} \;\leq\; &\left\{\mathbb{P}\left(\sum_{v=1}^{\infty}\lambda_v\xi_v^2\geq x\right)\right\}^{-1}\cdot\left\{\mathbb{P}\left(\sum_{v=1}^{\infty}\lambda_v\xi_v^2\geq\{(x-\epsilon_1)^{1/2}-\epsilon_2\}^2-\epsilon_3\right)\right. \\
&+C_3T^{5/2}\exp\left(-\frac{\sqrt{N}\epsilon_2}{C_4T^{5/2}}\right)+2\exp\left(-\frac{\epsilon_3}{C_5\sum_{v=K+1}^{\infty}\lambda_v}\right) \\
&\left.+\exp\left(-\frac{\epsilon_1}{C_9\sum_{v=T+1}^{\infty}\lambda_v}\right)\right\}. \qquad \text{(D.1)}
\end{aligned}$$

Our goal is now to show that the right-hand side of (D.1) converges to one by properly choosing $x, \epsilon_1, \epsilon_2, \epsilon_3, T$. To simplify the notation, we let $\zeta \overset{d}{=} \sum_{v=1}^{\infty}\lambda_v\xi_v^2$ and denote $\overline{F}_\zeta(x) = \mathbb{P}(\sum_{v=1}^{\infty}\lambda_v\xi_v^2\geq x)$. We also write the density function of ζ by $f_\zeta(x)$.

We start with the first term of (D.1). Write

$$\epsilon := x - \left\{(x-\epsilon_1)^{1/2} - \epsilon_2\right\}^2 - \epsilon_3.$$

Followed by Zolotarev (1961), we can approximate the survival function and density function of ζ as

$$\overline{F}_\zeta(x) = \frac{\kappa}{\Gamma(\mu_1/2)}\left(\frac{x}{2\lambda_1}\right)^{\mu_1/2-1}\exp\left(-\frac{x}{2\lambda_1}\right)\{1+o(1)\}$$
$$f_\zeta(x) = \frac{\kappa}{2\lambda_1\Gamma(\mu_1/2)}\left(\frac{x}{2\lambda_1}\right)^{\mu_1/2-1}\exp\left(-\frac{x}{2\lambda_1}\right)\{1+o(1)\}$$

for all $x > -\sum_{v=1}^{\infty}\lambda_v := -\Lambda$ that tends to infinity and $\kappa = \prod_{v=\mu_1+1}^{\infty}(1-\lambda_v/\lambda_1)^{-1/2}$. Then followed similarly by (A.13) of Drton et al. (2018), it is seen that there exists a constant x_0 such that for all $0 < \epsilon \le \lambda_1/2$,

$$\sup_{x\ge x_0}\left|\overline{F}_\zeta^{-1}(x)\cdot\max_{x'\in[x-\epsilon,x]} f_\zeta(x')\right| \le 2\lambda_1^{-1}.$$

Using this, the first term is bounded by

$$\begin{aligned}\frac{\mathbb{P}\left[\sum_{v=1}^{\infty}\lambda_v\xi_v^2 \ge \{(x-\epsilon_1)^{1/2}-\epsilon_2\}^2-\epsilon_3\right]}{\overline{F}_\zeta(x)} &\le \frac{\mathbb{P}\left(\sum_{v=1}^{\infty}\lambda_v\xi_v^2\ge x\right)}{\overline{F}_\zeta(x)} + \frac{\epsilon\cdot\max_{x'\in[x-\epsilon,x]} f_\zeta(x')}{\overline{F}_\zeta(x)}\\ &\le 1+2\epsilon\lambda_1^{-1}\end{aligned}$$

for all $x \ge x_0$. Next we shall choose $\epsilon_1, \epsilon_2, \epsilon_3$ decreasing in N so that $1+2\epsilon\lambda_1^{-1}$ converges uniformly to one for all $x \ge x_0$. Thus the upper bound of the first term converges to one uniformly over $x \ge x_0$. Hence, we only need to study the last three terms in (D.1) to finish the proof.

Let us first specify $T = \lfloor N^{(1-3\theta)/5}\rfloor$ where θ satisfies

$$\theta < \sup\left\{q\in[0,1/3) : \sum\nolimits_{v>\lfloor N^{(1-3q)/5}\rfloor}\lambda_v = O(N^{-q})\right\}.$$

Note that by the definition of θ, there exists a positive constant C_{10} such that $\sum_{v=T+1}^{\infty} \lambda_v \leq C_{10} N^{-\theta}$ for a sufficiently large N. Hence it now suffices to show that for all $x \in (0, o(N^{\theta}))$,

$$\begin{aligned}
&\left\{ \left(\frac{x}{2\lambda_1} \right)^{\mu_1/2-1} \exp\left(-\frac{x}{2\lambda_1} \right) \right\}^{-1} \exp\left(-\frac{\epsilon_1}{C_{11} N^{-\theta}} \right) \;\leq\; o(1), \\
&\left\{ \left(\frac{x}{2\lambda_1} \right)^{\mu_1/2-1} \exp\left(-\frac{x}{2\lambda_1} \right) \right\}^{-1} N^{(1-3\theta)/2} \exp\left(-\frac{\sqrt{N}\epsilon_2}{C_4 N^{(1-3\theta)/2}} \right) \;\leq\; o(1), \qquad \text{(D.2)} \\
&\left\{ \left(\frac{x}{2\lambda_1} \right)^{\mu_1/2-1} \exp\left(-\frac{x}{2\lambda_1} \right) \right\}^{-1} \exp\left(-\frac{\epsilon_3}{C_{12} N^{-\theta}} \right) \;\leq\; o(1).
\end{aligned}$$

For this purpose, we choose ϵ_1, ϵ_2 and ϵ_3 such that

$$\epsilon_1 = C_{11} N^{-\theta} \left(\frac{x}{2\lambda_1} + N^{\theta/2} \right), \quad \epsilon_2 = N^{-\theta/2}, \quad \epsilon_3 = C_{12} N^{-\theta} \left(\frac{x}{2\lambda_1} + N^{\theta/2} \right),$$

which tend to zero as $N \to \infty$ under $x \in (0, o(N^{\theta}))$. It is then straightforward to see that the three inequalities in (D.2) hold under the given setting. Consequently,

$$\frac{\mathbb{P}(n_1 n_2 \widehat{\mathcal{V}}_{12}^2 / N \geq x)}{\mathbb{P}\left(\sum_{v=1}^{\infty} \lambda_v \xi_v^2 \geq x\right)} \;\leq\; 1 + o(1).$$

The other direction follows similarly, which concludes

$$\frac{\mathbb{P}(n_1 n_2 \widehat{\mathcal{V}}_{12}^2 / N \geq x)}{\mathbb{P}\left(\sum_{v=1}^{\infty} \lambda_v \xi_v^2 \geq x\right)} \;=\; 1 + o(1)$$

uniformly over $x \in (0, o(N^{\theta}))$. This completes the proof.

D.1.2 Proof of Theorem 5.2

Continuing our discussion from Section 5.3.2, we apply Lemma 5.1.1 together with Theorem 5.1 to obtain the result. Specifically, we set

$$x = 4\lambda_1 \log K + \lambda_1(\mu_1 - 2) \log\log K + \lambda_1 y.$$

Then by the triangle inequality

$$\begin{aligned}
&\left| \mathbb{P}\big(n\widehat{\mathcal{V}}_{h,\max}^2/2 \leq x\big) - \exp\left\{ -\frac{2^{\mu_1/2-2}\kappa}{\Gamma(\mu_1/2)} \exp\left(-\frac{y}{2}\right) \right\} \right| \\
\leq\; &\left| \mathbb{P}\big(n\widehat{\mathcal{V}}_{h,\max}^2/2 \leq x\big) - \exp\left\{ -\frac{K(K-1)}{2} \mathbb{P}\big(n\widehat{\mathcal{V}}_{12}^2/2 > x\big) \right\} \right|
\end{aligned}$$

$$+\left|\exp\left\{-\frac{2^{\mu_1/2-2}\kappa}{\Gamma(\mu_1/2)}\exp\left(-\frac{y}{2}\right)\right\}-\exp\left\{-\frac{K(K-1)}{2}\mathbb{P}\big(n\widehat{\mathcal{V}}_{12}^2/2>x\big)\right\}\right|$$

$$=\ (I)+(II)\quad\text{(say)}.$$

By setting $\mathcal{I}=\{(i,j):1\le i<j\le K\}$ and $B_{u_{i,j}}=\{(k,l)\in\mathcal{I}:\text{Card}\{(k,l)\cap(i,j)\}\neq 0\}$ where $u_{i,j}:=(i,j)\in\mathcal{I}$ and Card$\{A\}$ denotes the cardinality of a set A, Lemma 5.1.1 yields $(I)\le b_1+b_2+b_3$. Here, in our setting,

$$b_1=\frac{K(K-1)(2K-3)}{2}\big\{\mathbb{P}\big(n\widehat{\mathcal{V}}_{12}^2/2>x\big)\big\}^2,$$

$$b_2=K(K-1)(K-2)\big\{\mathbb{P}\big(n\widehat{\mathcal{V}}_{12}^2/2>x\big)\big\}^2\quad\text{and}\quad b_3=0.$$

Therefore it is enough to verify that $\mathbb{P}(n\widehat{\mathcal{V}}_{12}^2/2>x)=O(K^{-2})$ under the given conditions. Then we have $(I)\to 0$ as $n,K\to\infty$.

In what follows, we prove $\mathbb{P}(n\widehat{\mathcal{V}}_{12}^2/2>x)=O(K^{-2})$ and $(II)\to 0$. First we apply Theorem 5.1 with $x\asymp 4\lambda_1\log K=o(n^\theta)$ to have

$$\frac{K(K-1)}{2}\mathbb{P}\big(n\widehat{\mathcal{V}}_{12}^2/2>x\big)=\frac{K(K-1)}{2}\mathbb{P}\bigg(\sum_{v=1}^{\infty}\lambda_v\xi_v^2>x\bigg)\{1+o(1)\}.$$

Using the tail approximation given by Zolotarev (1961) as $x\to\infty$:

$$\mathbb{P}\bigg(\sum_{v=1}^{\infty}\lambda_v\xi_v^2>x\bigg)=\frac{\kappa}{\Gamma(\mu_1/2)}\left(\frac{x}{2\lambda_1}\right)^{\mu_1/2-1}\exp\left(-\frac{x}{2\lambda_1}\right)\{1+o(1)\},$$

we have

$$\begin{aligned}\frac{K(K-1)}{2}\mathbb{P}\big(n\widehat{\mathcal{V}}_{12}^2/2>x\big)\ &=\ \frac{\kappa}{\Gamma(\mu_1/2)}\left(\frac{x}{2\lambda_1}\right)^{\mu_1/2-1}\exp\left(-\frac{x}{2\lambda_1}\right)\{1+o(1)\}\\ &=\ \exp\left\{-\frac{2^{\mu_1/2-2}\kappa}{\Gamma(\mu_1/2)}\exp\left(-\frac{y}{2}\right)\right\}\{1+o(1)\}.\end{aligned}$$

Therefore $\mathbb{P}(n\widehat{\mathcal{V}}_{12}^2/2>x)=O(K^{-2})$ and $(II)\to 0$ as $n,K\to\infty$, which completes the proof.

D.1.3 Proof of Theorem 5.5

Let us start by presenting some observations that are useful in the proof.

- **(O1).** From Lemma 5.4.1, we know that there exists a fixed constant $C_1 > 0$ such that

$$\max_{1\le k<l\le K} \left|\widehat{\mathcal{V}}_{kl} - \mathcal{V}_{kl}\right| \le C_1 \sqrt{\frac{B}{n_{\min}} \log\left(\frac{K}{\beta}\right)}$$

  with probability at least $1-\beta$.

- **(O2).** Let us define c_α such that

$$c_\alpha := \inf\left\{ t \in \mathbb{R} : \frac{1}{N!} \sum_{\boldsymbol{b}\in\mathcal{B}_N} \mathbb{1}\left(\widehat{\mathcal{V}}^{(\boldsymbol{b})}_{h,\max} \ge t\right) \le \alpha \right\}. \tag{D.3}$$

  From Theorem 5.4 and **(B2)**, there exists another fixed constant $C_2 > 0$ such that

$$c_\alpha \le C_2 \sqrt{\frac{B}{n_{\min}} \log\left(\frac{K}{\alpha}\right)}$$

  with probability one. Here we used the fact that $\widehat{\sigma}^2_K \le \max_{1\le i<j\le N} \widetilde{h}(Z_i, Z_j) \le B$. Thus the same inequality can be derived from (5.11) in Theorem 5.4 by replacing $\widehat{\sigma}^2_K$ with $\max_{1\le i<j\le N} \widetilde{h}(Z_i, Z_j)$, which is more efficient to compute.

- **(O3).** Based on the definition of c_α in (D.3), observe that the event $\{p_{\text{perm}} > \alpha\}$, which is equivalent to

$$\left\{ \frac{1}{N!} \sum_{\boldsymbol{b}\in\mathcal{B}_N} \mathbb{1}\left(\widehat{\mathcal{V}}^{(\boldsymbol{b})}_{h,\max} \ge \widehat{\mathcal{V}}_{h,\max}\right) > \alpha \right\},$$

  implies that $\{\widehat{\mathcal{V}}_{h,\max} \le c_\alpha\}$.

Having these observations in mind, let us define an event A_β such that

$$A_\beta = \left\{ \max_{1\le k<l\le K} \left|\widehat{\mathcal{V}}_{kl} - \mathcal{V}_{kl}\right| \le C_1 \sqrt{\frac{B}{n_{\min}} \log\left(\frac{K}{\beta}\right)} \right\}.$$

Then for sufficiently large $n_{\min}$, the type II error of the permutation test is bounded by

$$\begin{aligned} & \mathbb{P}\Big\{ p_{\text{perm}} > \alpha \Big\} \\ \overset{(i)}{\le}\ & \mathbb{P}\Big\{ \max_{1\le k<l\le K} \widehat{\mathcal{V}}_{kl} \le c_\alpha \Big\} \end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} \mathbb{P}\left\{ \max_{1\leq k<l\leq K} \widehat{\mathcal{V}}_{kl} \leq C_2\sqrt{\frac{B}{n_{\min}}\log\left(\frac{K}{\alpha}\right)} \right\} \\
&= \mathbb{P}\left\{ \max_{1\leq k<l\leq K} \widehat{\mathcal{V}}_{kl} \leq C_2\sqrt{\frac{B}{n_{\min}}\log\left(\frac{K}{\alpha}\right)},\ A_\beta \right\} + \mathbb{P}\left\{ \max_{1\leq k<l\leq K} \widehat{\mathcal{V}}_{kl} \leq C_2\sqrt{\frac{B}{n_{\min}}\log\left(\frac{K}{\alpha}\right)},\ A_\beta^c \right\} \\
&\overset{(iii)}{\leq} \mathbb{P}\left\{ \max_{1\leq k<l\leq K} \widehat{\mathcal{V}}_{kl} \leq C_2\sqrt{\frac{B}{n_{\min}}\log\left(\frac{K}{\alpha}\right)},\ A_\beta \right\} + \beta,
\end{aligned}
$$

where step (i) uses **(O3)**, step (ii) follows by **(O2)** and step (iii) uses **(O1)**. Furthermore, using the triangle inequality, we see that $\max_{1\leq k<l\leq K} \widehat{\mathcal{V}}_{kl} \geq \max_{1\leq k<l\leq K} \mathcal{V}_{kl} - \max_{1\leq k<l\leq K} \left|\widehat{\mathcal{V}}_{kl} - \mathcal{V}_{kl}\right|$. Also note that $\max_{1\leq k<l\leq K} \mathcal{V}_{kl} \geq b_N r_N^\star$ under the given condition. Thus

$$
\begin{aligned}
&\mathbb{P}\left\{ \max_{1\leq k<l\leq K} \widehat{\mathcal{V}}_{kl} \leq C_2\sqrt{\frac{B}{n_{\min}}\log\left(\frac{K}{\alpha}\right)},\ A_\beta \right\} \\
&\leq \mathbb{P}\left\{ b_N r_N^\star \leq C_1\sqrt{\frac{B}{n_{\min}}\log\left(\frac{K}{\beta}\right)} + C_2\sqrt{\frac{B}{n_{\min}}\log\left(\frac{K}{\alpha}\right)} \right\}.
\end{aligned}
$$

This gives an upper bound for the type II error that does not depend on $(P_1, \ldots, P_K)$. Since B is constant under **(B1)**, the upper bound goes to zero by taking e.g. $\beta = 1/b_N$. This completes the proof.

D.1.4 Proof of Corollary 5.5.1

First by the triangle inequality and Slutsky's argument,

$$
\begin{aligned}
\mathbb{P}(p_{\mathrm{MC}} > \alpha) &\leq \mathbb{P}(|p_{\mathrm{MC}} - p_{\mathrm{perm}}| + p_{\mathrm{perm}} > \alpha) \\
&\leq \mathbb{P}(|p_{\mathrm{MC}} - p_{\mathrm{perm}}| > \alpha/2) + \mathbb{P}(p_{\mathrm{perm}} > \alpha/2).
\end{aligned}
$$

Followed by Theorem 5.5, we have that

$$
\limsup_{n_{\min}\to\infty} \sup_{(P_1,\ldots,P_K)\in\mathcal{F}_h(b_N r_N^\star)} \mathbb{P}(p_{\mathrm{perm}} > \alpha/2) = 0.
$$

Therefore it suffices to control the first term. Let us write

$$
F(t) = \frac{1}{N!}\sum_{\boldsymbol{b}\in\mathcal{B}_N} \mathbb{1}\left(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b})} \leq t\right) \quad \text{and} \quad F_M(t) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}\left(\widehat{\mathcal{V}}_{h,\max}^{(\boldsymbol{b}_i')} \leq t\right).
$$

Then it can be shown that

$$|p_{\mathrm{perm}} - p_{\mathrm{MC}}| \;\leq\; \sup_{t\in\mathbb{R}} \left|F(t) - F_M(t)\right| + \frac{2}{M+1}.$$

Hence the first term is bounded by

$$\mathbb{P}(|p_{\mathrm{MC}} - p_{\mathrm{perm}}| > \alpha/2) \;\leq\; \mathbb{P}\left(\sup_{t\in\mathbb{R}} \left|F(t) - F_M(t)\right| > \frac{\alpha}{4}\right) + \mathbb{P}\left(\frac{2}{M+1} > \frac{\alpha}{4}\right).$$

Notice that by the DKW inequality (e.g. Massart, 1990),

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} \left|F(t) - F_M(t)\right| > \frac{\alpha}{4}\right) \leq 2e^{-M\alpha^2/8}.$$

Thus

$$\lim_{M\to\infty} \limsup_{n_{\min}\to\infty} \sup_{(P_1,\ldots,P_K)\in\mathcal{F}_h(b_N r_N^\star)} \mathbb{P}(|p_{\mathrm{MC}} - p_{\mathrm{perm}}| > \alpha/2) = 0,$$

which results in the conclusion.

D.1.5 Proof of Theorem 5.6

Motivated by Theorem 1 of Tolstikhin et al. (2017), we use discrete distributions to prove the result. Specifically, we choose two distinct points z_1, z_2 on $\mathbb{R}^d$ such that $\varphi(0) - \varphi(z_1 - z_2) \geq \kappa_1$. Consider the discrete distribution p_0 supported on the two points z_1, z_2 with probability $p_0(z_1) = 1/2$ and $p_0(z_2) = 1/2$. Consider another discrete distribution p_1 on the same support such that $p_1(z_1) = 1/2+\delta$ and $p_1(z_2) = 1/2-\delta$ where $\delta = br_N^\star/\sqrt{2\kappa_1}$ and b will be specified later. Then based on the translation invariant property of h, the MMD between p_0 and p_1 is calculated as

$$\mathcal{V}_h(p_0, p_1) = \delta\sqrt{2\{\varphi(0) - \varphi(z_1 - z_2)\}} \geq \delta\sqrt{2\kappa_1}. \tag{D.4}$$

See Tolstikhin et al. (2017) for details.

Next let k be a discrete random variable uniformly distributed on $\{1,\ldots,K\}$. Then we set $P_{1,k} = p_0\mathbb{1}(k \neq 1) + p_1\mathbb{1}(k = 1), \ldots, P_{K,k} = p_0\mathbb{1}(k \neq K) + p_1\mathbb{1}(k = K)$. Under this setting, it can be seen that $(P_{1,k},\ldots,P_{K,k}) \in \mathcal{F}_h(br_N^\star)$ using (D.4).

For each $k \in \{1,\ldots,K\}$, let q_k be the joint probability function of $X_{1,1},\ldots,X_{n_K,K}$ given by

$$q_k(x_{1,1},\ldots,x_{n_K,K}) = \prod_{i=1}^{n_1}\{p_0(x_{i,1})\mathbb{1}(k \neq 1) + p_1(x_{i,1})\mathbb{1}(k = 1)\}\times$$

$$\cdots \times \prod_{i=1}^{n_K} \{p_0(x_{i,K})\mathbb{1}(k \neq K) + p_1(x_{i,K})\mathbb{1}(k = K)\}.$$

Then we consider a mixture distribution given by $\overline{q}_{H_1} = \frac{1}{K}\sum_{k=1}^{K} q_k$. Also denote

$$q_{H_0}(x_{1,1}, \ldots, x_{n_K,K}) = \prod_{i=1}^{n_1} p_0(x_{i,1}) \times \cdots \times \prod_{i=1}^{n_K} p_0(x_{i,K}).$$

Then the likelihood ratio between $\overline{q}_{H_1}$ and q_{H_0} is

$$\begin{aligned} L_N = \frac{\overline{q}_{H_1}(X_{1,1}, \ldots, X_{n_K,K})}{q_{H_0}(X_{1,1}, \ldots, X_{n_K,K})} &= \frac{1}{K}\sum_{k=1}^{K}\prod_{i=1}^{n_k}\frac{p_1(X_{i,k})}{p_0(X_{i,k})} = \frac{1}{K}\sum_{k=1}^{K}\prod_{i=1}^{n_k}\frac{p_0(X_{i,k}) + \delta\gamma(X_{i,k})}{p_0(X_{i,k})} \\ &= \frac{1}{K}\sum_{k=1}^{K}\prod_{i=1}^{n_k}\{1 + 2\delta\gamma(X_{i,k})\}, \end{aligned}$$

where $\gamma(X_{i,k}) = \{\mathbb{1}(X_{i,k} = z_1) - \mathbb{1}(X_{i,k} = z_2)\}$. Moreover, the expected value of L_N^2 under H_0 is

$$\begin{aligned} \mathbb{E}_0(L_N^2) &= \frac{1}{K^2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\mathbb{E}_0\left[\prod_{i=1}^{n_k}\{1 + 2\delta\gamma(X_{i,k})\}\prod_{i=1}^{n'_k}\{1 + 2\delta\gamma(X_{i,k'})\}\right] \\ &= \frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}_0\left[\prod_{i=1}^{n_k}\{1 + 2\delta\gamma(X_{i,k})\}^2\right] + \frac{1}{K^2}\sum_{k\neq k'}^{K}\mathbb{E}_0\left[\prod_{i=1}^{n_k}\{1 + 2\delta\gamma(X_{i,k})\}\prod_{i=1}^{n'_k}\{1 + 2\delta\gamma(X_{i,k'})\}\right] \\ &= \frac{1}{K^2}\sum_{k=1}^{K}\prod_{i=1}^{n_k}\{1 + 4\delta^2\} + \frac{1}{K^2}\sum_{k\neq k'}^{K}\prod_{i=1}^{n_k}\prod_{i=1}^{n'_k}\{1\} \\ &\leq \frac{1}{K^2}\sum_{k=1}^{K}\exp\left(4n_k\delta^2\right) + \frac{K(K-1)}{K^2} \end{aligned}$$

where the last inequality uses $1 + x \leq e^x$ for all x. From the assumption **(B2)**, we know that there exists a fixed constant $C_3 > 0$ such that

$$\frac{1}{K^2}\sum_{k=1}^{K}\exp\left(4n_k\delta^2\right) \leq \frac{1}{K}\exp\left(C_3 n_{\min}\delta^2\right).$$

Finally, based on the standard χ^2 method for minimax testing (e.g. Baraud, 2002), it is enough to find a positive constant b such that $\delta = br_N^\star/\sqrt{2\kappa_1} < 1/2$ and $\mathbb{E}_0[L_N^2] \leq 1 + 4(1-\alpha-\zeta)^2$. Indeed, this holds for any $b < \min\{\sqrt{2\kappa_1/C_3}, \sqrt{\kappa_1}/(\sqrt{2}\kappa_2)\}$ for sufficiently large K, which completes the proof.

Appendix E

Appendix for Chapter 6

E.0.1 Proof of Lemma 1

To prove Lemma 1, we use the central limit theorem for a stationary α-mixing sequence.

Lemma E.0.1. *(Corollary 16.3.6 of Athreya and Lahiri, 2006) Let $\{U_n\}_{n\geq 1}$ be a sequence of stationary random variables with $E(U_1) = \mu_U \in \mathbb{R}$, $E|U_1|^{2+\delta} < \infty$ and $\sum_{i=1}^{\infty} \alpha_U(n)^{\delta/2+\delta} < \infty$ for some $\delta \in (0, \infty)$. Suppose that $\sigma^2_{\infty,U}$ is positive where*

$$\sigma^2_{\infty,U} = \lim_{n\to\infty} n^{-1}\text{var}\left(\sum_{i=1}^{n} U_i\right).$$

Then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} U_i - \mu_U\right)$$

converges to the normal distribution with mean zero and variance $\sigma^2_{\infty,U}$.

We first establish the multivariate central limit theorem for $d^{-1/2}\{(W_1^2 - \mu_1^2), \ldots, (W_M^2 - \mu_M^2)\}$ based on Lemma E.0.1 and apply the multivariate delta method to complete the proof. Let $\mathbb{S}^{M-1}$ be the M dimensional unit sphere defined by

$$\mathbb{S}^{M-1} = \left\{a \in \mathbb{R}^M : \sum_{i=1}^{M} a_i^2 = 1\right\}.$$

Then it is enough to show that for any $a = (a_1, \ldots, a_M)^\top \in \mathbb{S}^{M-1}$,

$$\frac{1}{\sqrt{d}}\{a_1(W_1^2 - \mu_1^2) + \cdots + a_M(W_M^2 - \mu_M^2)\} \tag{E.1}$$

converges to a normal distribution with mean zero and variance

$$\sigma^2_{\infty,a} = \lim_{d\to\infty} d^{-1}\mathrm{var}\{a_1(W_1^2 - \mu_1^2) + \cdots + a_M(W_M^2 - \mu_M^2)\}.$$

Here μ_i^2 is one of the values among $2d\sigma_x^2$, $2d\sigma_y^2$ and $d\sigma_x^2 + d\sigma_y^2 + d(\mu_x - \mu_y)^2$ depending on the class of Z_{i_2} and Z_{i_2} where $W_i^2 = ||Z_{i_1} - Z_{i_2}||_2^2$. In addition, the variance $\sigma^2_{\infty,a}$ is positive for all $a \in \mathbb{S}^{M-1}$ from the minimum eigenvalue condition (iii) in the main text.

Let $V_{1i} = (Z_{1i} - Z_{2i})^2, \ldots, V_{Mi} = (Z_{(N-1)i} - Z_{Ni})^2$ for $i = 1, \ldots, d$. Then we can rewrite

$$\frac{1}{\sqrt{d}}\{a_1(W_1^2 - \mu_1^2) + \cdots + a_M(W_M^2 - \mu_M^2)\} = \sqrt{d}\left\{\frac{1}{d}\sum_{i=1}^{d}\sum_{j=1}^{M} a_j V_{ji} - \sum_{j=1}^{M} a_j\mu_j^2\right\}.$$

Further let $V_{1:M,i,a} = \sum_{j=1}^{M} a_j V_{ji}$. Now we will show in two steps that $E|V_{1:M,1,a}|^{2+\delta} < \infty$ and $\sum_{r=1}^{\infty} \alpha_{V_{1:M,a}}(r)^{\delta/2+\delta} < \infty$ where $V_{1:M,a} = \{V_{1:M,i,a}\}_{i=1}^{\infty}$. Then apply Lemma E.0.1 to obtain the asymptotic multivariate normality of (E.1).

Step 1. In step 1, we will verify that $E|V_{1:M,1,a}|^{2+\delta} < \infty$. By Cauchy-Schwartz inequality and $||a||_2 = 1$, we have

$$\begin{aligned} E|V_{1:M,1,a}|^{2+\delta} &\le E\left(\sum_{j=1}^{M} V_{j1}^2\right)^{1+\delta/2} \\ &\le M^{\delta/2}\sum_{i=1}^{M} E|V_{j1}|^{2+\delta}, \end{aligned}$$

where the second inequality uses c_r-inequality. Therefore,

$$\begin{aligned} E|V_{1,a}|^{2+\delta} &\le M^{\delta/2}\left[E\{(Z_{11} - Z_{21})^{4+2\delta}\} + \cdots + E\{(Z_{(N-1)1} - Z_{N1})^{4+2\delta}\}\right] \\ &\le 2^{4+2\delta} M^{\delta/2+1} \max\{E|X_{11}|^{4+2\delta}, E|Y_{11}|^{4+2\delta}\}. \end{aligned}$$

Since M is fixed and $E|X_{11}|^{4+2\delta} < \infty$, $E|Y_{11}|^{4+2\delta} < \infty$ by the assumption (i) in the main text, the result follows.

Step 2. In step 2, we will show that $\sum_{r=1}^{\infty} \alpha_{V_{1:M,a}}(r)^{\delta/2+\delta} < \infty$. In this part of the proof, we proceed along the line with the proof of Theorem 2 in Li (2018). Recall that $\{Z_i\}_{i=1}^{N} = \{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$. Let $Z_{1:N,i} = (Z_{1i}, \ldots, Z_{Ni})^\top$ for $i = 1, \ldots, d$. Since the components of $Z_{1:N,i}$ are independent, the α-mixing

coefficient between two sigma fields $\sigma(\{Z_{1:N,i}\}_{i=1}^{k})$ and $\sigma(\{Z_{1:N,i}\}_{i=r+k}^{d})$ is bounded by

$$\begin{aligned}
& \alpha\{\sigma(\{Z_{1:N,i}\}_{i=1}^{k}), \sigma(\{Z_{1:N,i}\}_{i=r+k}^{d})\} \\
\leq & \sum_{j=1}^{N} \alpha\{\sigma(\{Z_{ji}\}_{i=1}^{k}), \sigma(\{Z_{ji}\}_{i=r+k}^{d})\} \\
\leq & \; m\alpha\{\sigma(\{X_{1i}\}_{i=1}^{k}), \sigma(\{X_{1i}\}_{i=r+k}^{d})\} + n\alpha\{\sigma(\{Y_{1i}\}_{i=1}^{k}), \sigma(\{Y_{1i}\}_{i=r+k}^{d})\}
\end{aligned}$$

for $k = 1, \ldots, d-r$. We now have $\sigma(V_{1:M,i,a}) \subseteq \sigma(Z_{1:N,i})$ since $V_{1:M,i,a}$ is a continuous function of $Z_{1:N,i}$. Therefore,

$$\begin{aligned}
& \alpha\{\sigma(\{V_{1:M,i,a}\}_{i=1}^{k}), \sigma(\{V_{1:M,i,a})\}_{i=r+k}^{d}\} \\
\leq & \; \alpha\{\sigma(\{Z_{1:N,i}\}_{i=1}^{k}), \sigma(\{Z_{1:N,i}\}_{i=r+k}^{d})\} \\
\leq & \; m\alpha\{\sigma(\{X_{1i}\}_{i=1}^{k}), \sigma(\{X_{1i}\}_{i=r+k}^{d})\} + n\alpha\{\sigma(\{Y_{1i}\}_{i=1}^{k}), \sigma(\{Y_{1i}\}_{i=r+k}^{d})\}.
\end{aligned}$$

This further implies that

$$\alpha_{V_{1:M,a}}(r) \leq m\alpha_X(r) + n\alpha_Y(r).$$

As a result, we obtain

$$\sum_{r=1}^{\infty} \alpha_{V_{1:M,a}}(r)^{\delta/2+\delta} \leq C_\delta \left\{ m^{\delta/2+\delta} \sum_{r=1}^{\infty} \alpha_X(r)^{\delta/2+\delta} + n^{\delta/2+\delta} \sum_{r=1}^{\infty} \alpha_Y(r)^{\delta/2+\delta} \right\},$$

where $C_\delta = \max\{1, 2^{3/2\delta-1}\}$. Since $\sum_{r=1}^{\infty} \alpha_X(r)^{\delta/2+\delta} < \infty$, $\sum_{r=1}^{\infty} \alpha_Y(r)^{\delta/2+\delta} < \infty$ from the assumption (ii) in the main text and m, n are fixed, the result follows.

Finally, we conclude that $d^{-1/2}\{(W_1^2 - \mu_1^2), \ldots, (W_M^2 - \mu_M^2)\}$ converges to a multivariate normal distribution with mean zero and a positive definite covariance matrix and complete the proof by applying the multivariate delta method.

E.0.2 Proof of Theorem 1

We prove the given statement for each test as follows.

- Friedman and Rafsky' test: recall that we reject the null for a small value of $T_{m,n}^{\mathrm{FR}}$, and hence, it is enough to verify that the maximum of $T_{m,n}^{\mathrm{FR}}$ can be observed with nonzero probability under

the alternative. The maximum of $T_{m,n}^{\mathrm{FR}}$ is obtained as $T_{m,n}^{\mathrm{FR}} = N$ when any edge on the minimal spanning tree connects two observations from the different distributions. This maximum value can be observed, for example, when the maximum of between class distances is less than the minimum of within class distances. The probability of the described event converges to some positive value $\delta > 0$ as $d \to \infty$, due to Lemma 1 under the given conditions. Let us denote the cut-off value of $T_{m,n}^{\mathrm{FR}}$ by c_α for some $\alpha \in (0,1)$. Then the asymptotic power of the test is calculated by $\lim_{d\to\infty} \mathrm{pr}_{H_1}(T_{m,n}^{\mathrm{FR}} < c_\alpha) \leq 1 - \lim_{d\to\infty} \mathrm{pr}_{H_1}(T_{m,n}^{\mathrm{FR}} = N) < 1 - \delta$ as $d \to \infty$. This shows that the Friedman and Rafsky' test cannot be consistent for any $\alpha \in (0,1)$.

- Nearest neighbor test: the nearest neighbor test rejects the null for a large value of $T_{m,n}^{\mathrm{NN}}$. Again, it is enough to show that the minimum value of $T_{m,n}^{\mathrm{NN}}$ has a positive probability under the alternative. This can be shown by taking the same example considered in Friedman and Rafsky's test, which in turn gives the result.

- Baringhaus and Franz's test: for Baringhaus and Franz's test, we reject the null for a large value of $T_{m,n}^{\mathrm{BF}}$. Since $T_{m,n}^{\mathrm{BF}}$ is a linear combination of $(W_1, \ldots, W_M)$, we have the asymptotic normality of $T_{m,n}^{\mathrm{BF}}$ under both the null and alternative hypotheses. Let us denote the asymptotic variance of $T_{m,n}^{\mathrm{BF}}$ by $\sigma_{0,\mathrm{BF}}^2$ and $\sigma_{1,\mathrm{BF}}^2$ under the null and alternative, respectively. Since the asymptotic means of $T_{m,n}^{\mathrm{BF}}$ are identical under both null and alternative hypotheses, the power of $T_{m,n}^{\mathrm{BF}}$ converges to

$$\Phi\left(-\frac{\sigma_{0,\mathrm{BF}}}{\sigma_{1,\mathrm{BF}}} z_\alpha\right), \tag{E.2}$$

  where Φ is the standard normal distribution function and z_α is the corresponding upper α quantile. Therefore, for any $\alpha \in (0,1)$, Baringhaus and Franz's test cannot be consistent.

E.0.3 Proof of Theorem 2

Under the given assumptions, the usual multivariate central limit theorem shows that there exists a covariance matrix Σ such that

$$\frac{1}{\sqrt{d}}\{(W_1^2, \ldots, W_M^2)^\top - (\mu_1^2, \ldots, \mu_M^2)^\top\}$$

converges to $N_d(0, \Sigma)$. Since $E|X_{11}|^p = E|Y_{11}|^p < \infty$ for $p = 1, 2, 3, 4$, it is seen that the mean vector $(\mu_1^2, \ldots, \mu_M^2)^\top$ and the covariance matrix Σ are the same under both null and alternative hypotheses. As a

result, the asymptotic null and alternative distributions of $g(W_1, \ldots, W_M) - g(\mu_1, \ldots, \mu_M)$ are the same by the delta method. Therefore, the result follows.

E.0.4 Proof of Lemma 2

Based on Lemma 2.1 of Baringhaus and Franz (2004), it is seen that

$$\begin{aligned} E|X_{11} - X_{21}| &= 2\int_{-\infty}^{\infty} F_{X_{11}}(t)\,(1 - F_{X_{11}}(t))\,dt, \\ E|Y_{11} - Y_{21}| &= 2\int_{-\infty}^{\infty} G_{Y_{11}}(t)\,(1 - G_{Y_{11}}(t))\,dt, \\ E|X_{11} - Y_{11}| &= \int_{-\infty}^{\infty} G_{Y_{11}}(t)\,(1 - F_{X_{11}}(t))\,dt + \int_{-\infty}^{\infty} F_{X_{11}}(t)\,(1 - G_{Y_{11}}(t))\,dt. \end{aligned} \tag{E.3}$$

Therefore, we have $\gamma_x = E|X_{11} - X_{21}|$, $\gamma_y = E|Y_{11} - Y_{21}|$ and $\gamma_{xy} = E|X_{11} - Y_{11}|$. Suppose that we are at the null hypothesis. Then it is trivial to see that $\gamma_x = \gamma_y = \gamma_{xy}$. Now suppose that $\gamma_x = \gamma_y = \gamma_{xy}$, which in turn implies $2\gamma_{xy} = \gamma_x + \gamma_y$. According to Lemma 2.2 of Baringhaus and Franz (2004), it is always true that $2\gamma_{xy} \geq \gamma_x + \gamma_y$ and the equality holds if and only if $F_{X_{11}} = G_{Y_{11}}$. Hence, we conclude that $\gamma_x = \gamma_y = \gamma_{xy}$ if and only if the marginal distributions of X and Y are the same.

Consider sequences $D_x = \{|X_{1i} - X_{2i}|\}_{i=1}^{\infty}$, $D_y = \{|Y_{1i} - Y_{2i}|\}_{i=1}^{\infty}$ and $D_{xy} = \{|X_{1i} - Y_{1i}|\}_{i=1}^{\infty}$. Then similarly to step 2 of the proof of Lemma 1, it is seen that

$$\sum_{r=1}^{\infty} \alpha_{D_x}(r)^{\delta/2+\delta} < \infty, \quad \sum_{r=1}^{\infty} \alpha_{D_y}(r)^{\delta/2+\delta} < \infty, \quad \sum_{r=1}^{\infty} \alpha_{D_{xy}}(r)^{\delta/2+\delta} < \infty.$$

Now based on Proposition 16.3.1 of Athreya and Lahiri (2006) together with Chebyshev's inequality, we conclude that $d^{-1}||X_1 - X_2||_1$, $d^{-1}||Y_1 - Y_2||_1$ and $d^{-1}||X_1 - Y_1||_1$ converge in probability to γ_x, γ_y and γ_{xy}, respectively.

E.0.5 Details of Example 1

Since X_1 and X_2 are independent and have the standard normal distribution, $|X_1 - X_2|$ has the folded normal with parameters $(\mu = 0, \sigma^2 = 2)$. Hence from (E.3), we have

$$\gamma_x = E|X_1 - X_2| = \frac{2}{\sqrt{\pi}}.$$

Let U have the normal distribution with mean μ_1 and variance σ_1^2. Similarly, let V have the normal distribution with mean μ_2 and variance σ_2^2. In this case, the expected value of $|Y_1 - Y_2|$ can be calculated by

$$E|Y_1 - Y_2| = \lambda^2 E|U_1 - U_2| + 2\lambda(1-\lambda)E|U_1 - V_1| + (1-\lambda)^2 E|V_1 - V_2|.$$

Since $|U_1 - U_2|$ has the folded normal distribution with parameters $(\mu = 0, \sigma^2 = 2\sigma_1^2)$, we obtain $E|U_1 - U_2| = f(0, 2\sigma_1^2)$. Similarly, we have $E|U_1 - V_1| = f(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$ and $E|V_1 - V_2| = f(0, 2\sigma_2^2)$. This gives the expression for γ_y. Lastly, we have

$$E|X_1 - Y_1| = \lambda E|X_1 - U_1| + (1-\lambda)E|X_1 - V_1|,$$

which gives the expression for γ_{xy}.

E.0.6 Proof of Theorem 3

(i) The first result is a direct consequence of Theorem 2 in Biswas et al. (2014), and so is omitted.

(ii) When $F_X \neq G_Y$, the nearest neighbor test statistic $T_{m,n}^{\mathrm{NN}}$ converges in probability to one as $d \to \infty$. Since the maximum value of $T_{m,n}^{\mathrm{NN}}$ is equal to one, it is enough to show that the α level critical value of the permutation test is less than one.

Without loss of generality, assume that $m \leq n$. We claim that there are $1 + n!/\{m!(m-n)!\}$ permutations of the class labels that can potentially achieve the maximum under the given conditions. If this is the case, then the α level critical value of the permutation test becomes less than one by choosing $\alpha > [1 + n!/\{m!(m-n)!\}]/\{N!/(m!n!)\}$. Hence the result follows. We write the original samples $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ by $\{(Z_1, L_1), \ldots, (Z_m, L_m), (Z_{m+1}, L_{m+1}), \ldots, (Z_{m+n}, L_{m+n})\}$ where $Z_i \in \mathbb{R}^d$ and $(L_1, \ldots, L_N) = (1, \ldots, 1, 0, \ldots, 0)$. By the assumption that $\gamma_{xy} > \max\{\gamma_x, \gamma_y\}$ and $k < m$, there will be no connected edge between $\mathcal{Z}_m = \{Z_1, \ldots, Z_m\}$ and $\mathcal{Z}_n = \{Z_{m+1}, \ldots, Z_{m+n}\}$ in the k-nearest neighbor graph as $d \to \infty$. This means that the k nearest neighbor of any $Z_i \in \mathcal{Z}_m$ is an element of $\mathcal{Z}_m$. Now assume that the first m permuted labels are neither $(L_1, \ldots, L_m) = (1, \ldots, 1)$ nor $(L_1, \ldots, L_m) = (0, \ldots, 0)$. Then by the assumption that $k \geq m/2$, there exists at least one $Z_i \in \mathcal{Z}_m$ such that

$$\sum_{r=1}^{k} I_i(r) < k$$

where $I_i(r)$ is the indicator variable equal to one if and only if Z_i and its rth nearest neighbor have the same class label. This implies that $T_{m,n}^{\mathrm{NN}} < 1$. Now the number of permutations which have either $(L_1, \ldots, L_m) =$

$(1,\ldots,1)$ or $(L_1,\ldots,L_m)=(0,\ldots,0)$ is $1+n!/\{m!(m-n)!\}$ out of $N!/(m!n!)$ total permutations. Therefore the result follows.

(*iii*) For Baringhaus and Franz's test, let us write

$$D_{xy,m,n}=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}d^{-1}||X_i-Y_j||_1,\quad D_{x,m}=\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}d^{-1}||X_i-X_j||_1,$$

$$D_{y,n}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}d^{-1}||Y_i-Y_j||_1.$$

Then from Lemma 2, we see that $D_{xy,m,n}$ converges in probability to γ_{xy}. Similarly, we have that $D_{x,m}$ and $D_{y,n}$ converge in probability to $\{(m-1)/m\}\gamma_x$ and $\{(n-1)/n\}\gamma_y$, respectively. Consequently, $d^{-1}T^{\mathrm{BF}}_{m,n}$ converges to $\gamma_{xy}-\{(m-1)/(2m)\}\gamma_x-\{(n-1)/(2n)\}\gamma_y$ in probability.

Suppose that $\gamma_{xy}>\max\{\gamma_x,\gamma_y\}$. Then it can be seen that $d^{-1}T^{\mathrm{BF}}_{m,n}$ becomes the maximum value among the $N!/\{m!n!\}$ permutation statistics as d tends to infinity. If $m\neq n$, there exists one unique permutation that returns the maximum value of $d^{-1}T^{\mathrm{BF}}_{m,n}$. On the other hand, if $m=n$, there exist two permutations (the original permutation and its complement) that return the maximum value of $d^{-1}T^{\mathrm{BF}}_{m,n}$. Hence, if $\alpha>(m!n!)/N!$ for $m\neq n$ and if $\alpha>2(m!n!)/N!$ when $m=n$, the power of Baringhaus and Franz's test converges to one as $d\to\infty$.

E.0.7 Additional Simulations

This section provides additional simulation results. Let A be an autocorrelation matrix where $[A]_{i,i}=1$ and $[A]_{i,j}=0.2^{|i-j|}$ for $i\neq j$. Then we transform random vectors by $\widetilde{X}=AX$ and $\widetilde{Y}=AY$ where $X\sim F_X$ and $Y\sim G_Y$. We consider the same location, scale and kurtosis examples for F_X and G_Y in Section 5 of the main text, but now the components of $\widetilde{X}$ and $\widetilde{Y}$ are weakly dependent. The results are presented in Table E.1. As in the independent case in the main text, the Manhattan-based tests outperform the Euclidean-based tests against the kurtosis alternative, whereas they have similar performance to the Euclidean-based tests against the location and scale alternatives.

Table E.1: Empirical power of the tests over different dimensions at significance level $\alpha = 0.05$ and $m = n = 20$ when covariates are weakly dependent.

		Location			*Scale*			*Kurtosis*		
	d	100	500	1000	100	500	1000	100	500	1000
$T^{\mathrm{BF}}_{m,n}$	*Manhattan*	0.950	1.000	1.000	0.087	0.160	0.199	**0.320**	**0.893**	**0.994**
	Euclidean	0.961	1.000	1.000	0.086	0.156	0.209	0.061	0.044	0.051
$T^{\mathrm{NN}}_{m,n}$	*Manhattan*	0.697	0.997	1.000	0.067	0.012	0.002	**0.273**	**0.615**	**0.805**
	Euclidean	0.712	1.000	1.000	0.074	0.012	0.001	0.063	0.050	0.062
$T^{\mathrm{FR}}_{m,n}$	*Manhattan*	0.533	0.967	0.999	0.068	0.001	0.000	**0.240**	**0.430**	**0.573**
	Euclidean	0.557	0.971	1.000	0.062	0.001	0.000	0.078	0.068	0.075
$T^{\mathrm{MGB}}_{m,n}$	*Manhattan*	0.170	0.786	0.985	0.798	1.000	1.000	**0.101**	**0.385**	**0.686**
	Euclidean	0.191	0.807	0.984	0.799	1.000	1.000	0.071	0.070	0.080
$T^{\mathrm{CF}}_{m,n}$	*Manhattan*	0.178	0.643	0.919	0.671	0.998	1.000	**0.135**	**0.358**	**0.613**
	Euclidean	0.181	0.655	0.922	0.678	0.997	1.000	0.099	0.090	0.079

Appendix F

Appendix for Chapter 7

F.1 Outline

This chapter is organized as follows. In Section F.2, we discuss some open problems, raised by our main results. Section F.3.1 contains some lemmas that will prove useful in many of the proofs. In Section F.3.2, we provide the proof of Proposition 7.1, which shows the asymptotic expression for the minimax power. Section F.3.3 presents the proof of Theorem 7.1, which demonstrates the optimality of Hotelling's T^2 test when $d/n \to 0$. Section F.3.4 focuses on Proposition 7.2 and proves the asymptotic normality of W_A. In Section F.3.5, Theorem 7.2 and Theorem 7.4 are proved, verifying the asymptotic normality of $W_A^\dagger$ and the asymptotic power of the naive Bayes classifier test. Section F.3.6 proves Lemma F.0.4. By building on some moment expressions for (scaled) inverse chi-square random variables in Section F.3.7, we provide the proof of Lemma F.0.5 in Section F.3.8. Section F.3.9 provides the proof of Theorem 7.5, which is an extension of our main result to elliptical distributions. In Section F.3.10 and Section F.3.11, we prove the type-1 error control and consistency result of the asymptotic test and the permutation test, respectively. Lastly, some simulation results on sample-splitting ratio are presented in Section F.4.

F.2 Open problems

Here we discuss how our results may be extended to a larger context while we leave a detailed analysis to future work. Four open problems that we first highlight are as follows:

- The most obvious open problem is to extend our power guarantees, and most other published ones in the high-dimensional two-sample testing literature, to be uniform over an entire class of alternative distributions rather than just holding pointwise. Viewing our proofs in this material, uniform control of the relevant error terms seems extremely challenging.

- Determining whether our minimax lower bound can be achieved by any test when $d = O(n)$, or if tighter lower bounds can be proved, is an important open problem. Of course, we have settled this problem for $d = o(n)$ in Section 7.4 even from the perspective of uniformity.

- Given that the focus of this study is mainly on Fisher's LDA classifier and its variants, there is a possibility that some other linear discrimination rules (e.g. via empirical risk minimization) may achieve optimal power.

- Beyond the consistency result, proving that one can achieve the same non-trivial power as the asymptotic tests using permutations seems like an interesting open problem.

From the perspective of the title of the current paper, we provide other four natural directions for future explorations.

Leave-one-out accuracy Another natural estimator for accuracy, as an alternative to sample-splitting, is a leave-one-out estimator $\widehat{E}^L$, defined as $\widehat{E}^L \stackrel{\text{def}}{=} (\widehat{E}^L_0 + \widehat{E}^L_1)/2$, with

$$\begin{aligned} \widehat{E}^L_0 &\stackrel{\text{def}}{=} \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbb{I}\Big[\mathrm{LDA}_{n_0\setminus i, n_1}(X_i) = 1\Big], \\ \widehat{E}^L_1 &\stackrel{\text{def}}{=} \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{I}\Big[\mathrm{LDA}_{n_0, n_1\setminus i}(Y_i) = 0\Big], \end{aligned} \tag{F.1}$$

where $\mathrm{LDA}_{n_0\setminus i, n_1}$ (or $\mathrm{LDA}_{n_0, n_1\setminus i}$) denotes the LDA classifier using all points except X_i (or Y_i).

Ensemble accuracy The sample-splitting estimator $\widehat{E}^S$ in (7.5) is based on an arbitrary split in training and test sets. Hence the resulting test is potentially unstable depending on the result of sample splitting. This issue can be simply overcome by considering all possible splits. Let $\sigma := \{\sigma(1), \ldots, \sigma(n_{0,\mathrm{tr}})\}$ be a subset of $\{1, \ldots, n_0\}$ drawn without replacement. Similarly, let $\sigma' := \{\sigma'(1), \ldots, \sigma'(n_{1,\mathrm{tr}})\}$ be a subset of $\{1, \ldots, n_1\}$ drawn without replacement. By setting $\{X_{\sigma(1)}, \ldots, X_{\sigma(n_{0,\mathrm{tr}})}\} \cup \{Y_{\sigma'(1)}, \ldots, Y_{\sigma'(n_{1,\mathrm{tr}})}\}$ as the training set and the remaining as the test set, one can calculate $\widehat{E}^S := \widehat{E}^S(\sigma, \sigma')$. The ensemble estimator is then defined by

$$\widehat{E}^{Ens} = \frac{1}{\binom{n_0}{n_{0,\mathrm{tr}}}\binom{n_1}{n_{1,\mathrm{tr}}}} \sum_{\sigma}\sum_{\sigma'} \widehat{E}^S(\sigma, \sigma'),$$

where the first sum is taken over all possible subsets of size $n_{0,\mathrm{tr}}$ from $\{1, \ldots, n_0\}$ and the second sum is taken over all possible subsets of size $n_{1,\mathrm{tr}}$ from $\{1, \ldots, n_1\}$. Although it looks similar to the U-statistic

considered in Hediger et al. (2019), the ensemble estimator differs from theirs by allowing $\widehat{E}^S(\sigma, \sigma')$ to be a function of the entire dataset rather than a subset. Hence the proposed one uses the dataset more efficiently.

Resubstitution accuracy Since leave-one-out estimators and ensemble estimators are computationally intensive, one might be tempted to use the training data itself to test the classifier. This resubstitution error would be defined as $\widehat{E}^R \stackrel{\text{def}}{=} (\widehat{E}_0^R + \widehat{E}_1^R)/2$, with

$$\begin{aligned} \widehat{E}_0^R &\stackrel{\text{def}}{=} \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbb{I}\Big[\text{LDA}_{n_0,n_1}(X_i) = 1\Big], \\ \widehat{E}_1^R &\stackrel{\text{def}}{=} \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{I}\Big[\text{LDA}_{n_0,n_1}(Y_i) = 0\Big], \end{aligned} \tag{F.2}$$

where we first train on all the data and then test on all the data. Of course such an estimate would be overoptimistic, and would be scorned upon as an estimate of the true accuracy E of the classifier. However, one might hope that the null distribution or permutation distribution would be similarly optimistically biased (instead of being centered around a half), thus nullifying the optimistic bias of $\widehat{E}^R$. From simulation studies, we observed that the accuracy test based on the resubstitution error performs slightly better than the test based on sample splitting in low-dimensional scenarios but overall performs similarly (e.g. Figure 7.3). It will be interesting to theoretically justify the asymptotic behavior of resubstitution accuracy (and also leave-one-out and ensemble accuracy) and see whether the resulting test is also minimax rate optimal.

Non-linear classification Another natural setting is that of nonlinear classification. An examination of the test statistics used (Hotelling and its variants) shows that they are closely related to the statistics based on the kernel Maximum Mean Discrepancy (Gretton et al., 2012) and the kernel FDA (Eric et al., 2008), when specifically instantiated with the linear kernel. Similarly, for classification, a kernelized LDA (Mika et al., 1999) specializes to Fisher's LDA when the linear kernel is employed.

Given the parallels observed, and given that a kernel classifier or two-sample test is effectively a linear method in a higher dimensional space, one might naturally conjecture that the spirit of the results of this paper can be extended to such kernelized nonlinear settings as well. As mentioned before, very recent progress has been made by Hediger et al. (2019) for random forests (but not in the high dimensional setting).

The use of neural network type classifiers for classifier-based testing on structured data is certainly an interesting direction, though precise theoretical characterizations, such as the ones provided in this paper, seem unlikely given our current understanding.

F.3 Technical proofs

F.3.1 Supporting lemmas

Before we present the detailed proofs of all our results, we collect some supporting lemmas. The first lemma provides the mean and variance of a quadratic form of Gaussian random vectors.

Lemma F.0.1 (Chapter 5.2 in Rencher and Schaalje (2008)). *Suppose that Z has a multivariate Gaussian distribution with mean μ and covariance Σ. Then, we have*

$$\mathbb{E}[Z^\top \Lambda Z] = \mathrm{tr}[\Lambda\Sigma] + \mu^\top \Lambda \mu \text{ and}$$

$$\mathbb{V}[Z^\top \Lambda Z] = 2\mathrm{tr}[\Lambda\Sigma\Lambda\Sigma] + 4\mu^\top \Lambda\Sigma\Lambda\mu.$$

Next we present the Berry-Esseen theorem for non-identically distributed summands, which will be used to prove Proposition 7.2.

Lemma F.0.2 (Berry-Esseen theorem, Berry (1941)). *Let $X_1, X_2, \ldots,$ be independent random variables with $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = \sigma_i^2 > 0$ and $\mathbb{E}[|X_i|^3] = \rho_i < \infty$. Define $S_n = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$ and let $F_n(\cdot)$ be its CDF. Then there exists a constant $C_1 > 0$ such that*

$$\sup_{t\in\mathbb{R}} |F_n(t) - \Phi(t)| \leq \frac{C_1}{\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}} \max_{1\leq i\leq n} \frac{\rho_i}{\sigma_i^2}.$$

The following lemma bounds the trace of a product of two matrices in terms of their eigenvalues.

Lemma F.0.3 (Fan's inequality, page 10 of Borwein and Lewis (2010)). *For any symmetric matrices $A, B \in \mathbb{R}^{d\times d}$, we have $\mathrm{tr}(AB) \leq \sum_{i=1}^d \lambda_i(A)\lambda_i(B)$.*

Before stating the next two lemmas, let us recall some notation from the main text. First $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ are the errors of the generalized LDA conditional on the input data. These can be written as $\mathcal{E}_{0,A} = \Phi(V_{0,A}/\sqrt{U_A})$ and $\mathcal{E}_{1,A} = \Phi(V_{1,A}/\sqrt{U_A})$ where $V_{0,A} = \widehat{\delta}^\top A(\mu_0 - \widehat{\mu}_{\mathrm{pool}})$, $V_{1,A} = \widehat{\delta}^\top A(\widehat{\mu}_{\mathrm{pool}} - \mu_1)$ and $U_A = \widehat{\delta}^\top A\Sigma A\widehat{\delta}$. Further recall $\Psi_{A,n,d} = -\delta^\top A\delta/2$, $\Lambda_{A,n,d} = \delta^\top A\Sigma A\delta + (1/n_{0,\mathrm{tr}} + 1/n_{1,\mathrm{tr}})\,\mathrm{tr}\{(A\Sigma)^2\}$ and $\Xi_{A,n,d} = (n_{0,\mathrm{tr}}^{-1} - n_{1,\mathrm{tr}}^{-1})\mathrm{tr}(A\Sigma)/2$.

The following lemma presents approximations of $\mathcal{E}_{0,A}$, $\mathcal{E}_{1,A}$ and $\mathcal{E}_{0,A} + \mathcal{E}_{1,A}$, which plays a key role in proving Theorem 7.2.

Lemma F.0.4. *Under the assumptions **(A1)**–**(A6)**, $\mathcal{E}_{0,A}$, $\mathcal{E}_{1,A}$ and $\mathcal{E}_{0,A} + \mathcal{E}_{1,A}$ are*

$$\mathcal{E}_{0,A} \;=\; \Phi\left(\frac{\Psi_{A,n,d} + \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + O_P\left(n^{-1/2}\right),$$

$$\mathcal{E}_{1,A} = \Phi\left(\frac{\Psi_{A,n,d} - \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + O_P\left(n^{-1/2}\right) \text{ and}$$

$$\mathcal{E}_{0,A} + \mathcal{E}_{1,A} = \Phi\left(\frac{\Psi_{A,n,d} + \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + \Phi\left(\frac{\Psi_{A,n,d} - \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + o_P\left(n^{-1/2}\right).$$

Furthermore, when $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$,

$$\mathcal{E}_{0,A} + \mathcal{E}_{1,A} = 2\Phi\left(\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + O_P(n^{-3/4}).$$

One thing to notice from the above lemma is that the sum of $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ converges faster than the individual components because the higher order error terms cancel out in the sum. This critical phenomenon allows us to replace $\mathcal{E}_{0,A}/2 + \mathcal{E}_{1,A}/2$ in W_A with a non-random counterpart. The proof of Lemma F.0.4 can be found in Section F.3.6.

The following lemma is similar to Lemma F.0.4 but by replacing a non-random matrix A with random diagonal matrix $\widehat{D}^{-1} = \mathrm{diag}(\widehat{\Sigma})^{-1}$. This lemma will be used to prove Theorem 7.4 where we present the power of the naive Bayes classifier test.

Lemma F.0.5. *Assume that* $n_0 = n_1$, $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$ *and* $n_{\mathrm{tr}} = n_{\mathrm{te}}$. *Then under the assumptions* ***(A1)***, ***(A2)*** *and* ***(A5)***, $\mathcal{E}_{0,\widehat{D}^{-1}}$, $\mathcal{E}_{1,\widehat{D}^{-1}}$ *and* $\mathcal{E}_{0,\widehat{D}^{-1}} + \mathcal{E}_{1,\widehat{D}^{-1}}$ *are*

$$\mathcal{E}_{0,\widehat{D}^{-1}} = \Phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right) + O_P\left(n^{-1/2}\right),$$

$$\mathcal{E}_{1,\widehat{D}^{-1}} = \Phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right) + O_P\left(n^{-1/2}\right) \text{ and}$$

$$\mathcal{E}_{0,\widehat{D}^{-1}} + \mathcal{E}_{1,\widehat{D}^{-1}} = 2\Phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right) + O_P\left(n^{-3/4}\right).$$

As in Lemma F.0.4, due to the cancellation of higher order error terms, we observe that the sum of $\mathcal{E}_{0,\widehat{D}^{-1}}$ and $\mathcal{E}_{1,\widehat{D}^{-1}}$ converges faster than either individual component. The proof of Lemma F.0.5 can be found in Section F.3.8.

We now have all the results in place to prove the main results in the paper.

F.3.2 Proof of Proposition 7.1 (minimax lower bound)

We begin by recalling the result by Luschgy (1982) in (7.6), which implies that to derive a bound on the minimax power, one only needs to analyze the power of the oracle Hotelling's procedure φ_H^* with known

Σ. Next, note that $\frac{n_0 n_1}{n_0+n_1}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \Sigma^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1)$ has a noncentral chi-square distribution with d degrees of freedom and noncentrality parameter $\frac{n_0 n_1}{n_0+n_1}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)$. Using the monotonicity of the distribution function of a noncentral chi-square random variable in its non-centrality parameter, it can thus be seen that

$$\sup_{\varphi_\alpha \in \mathcal{T}_\alpha} \inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_\alpha] = \inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_H^*] = \mathbb{P}\left(\sum_{i=1}^d Z_i^2 \geq c_{\alpha,d}\right), \tag{F.3}$$

where $Z_i \overset{i.i.d.}{\sim} N(\rho_n, 1)$ and $\rho_n \overset{\text{def}}{=} \sqrt{\frac{n_0 n_1}{n_0+n_1}}\rho$. Note that the right-hand side of (F.3) rearranges to

$$\mathbb{P}\left(\frac{\sum_{i=1}^d Z_i^2 - d - \rho_n^2}{\sqrt{2(d+2\rho_n^2)}} \geq \frac{c_{\alpha,d} - d}{\sqrt{2d}} \frac{\sqrt{2d}}{\sqrt{2(d+2\rho_n^2)}} - \frac{\rho_n^2}{\sqrt{2(d+2\rho_n^2)}}\right).$$

By Lyapunov's central limit theorem, we know that

$$\frac{\sum_{i=1}^d Z_i^2 - d - \rho_n^2}{\sqrt{2(d+2\rho_n^2)}} \xrightarrow{d} N(0,1) \quad \text{and} \quad \frac{c_{\alpha,d} - d}{\sqrt{2d}} \to z_\alpha, \tag{F.4}$$

using which the statement of Proposition 7.1 immediately follows.

F.3.3 Proof of Theorem 7.1 (optimality of Hotelling's T^2 test)

We first describe a couple of preliminaries and then prove the main theorem.

Preliminaries Under the Gaussian setting, it is well-known (**?**) that

$$\frac{n_0 + n_1 - 1 - d}{d(n_0 + n_1 - 2)} \frac{n_0 n_1}{n_0 + n_1} T_H \sim F(d, n-1-d; \rho_n^2),$$

where $F(d, n-1-d; \rho_n^2)$ has the non-central F-distribution with noncentrality parameter $\rho_n^2 = \frac{n_0 n_1}{n_0+n_1}(\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)$ and d and $n-1-d$ degrees of freedom. Let $\chi_d^2(\rho_n^2)$ be a noncentral chi-square random variable with noncentrality parameter ρ_n^2 and d degrees of freedom and write $\chi_{n-1-d}^2(0) = \chi_{n-1-d}^2$ for simplicity. Using the monotonicity of the distribution function of a noncentral F random variable in its non-centrality parameter, it can be seen that

$$\inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_H] = \mathbb{P}\{F(d, n-1-d; \rho_n^2) \geq q_{\alpha,n,d}\}.$$

Hence it is enough to study the asymptotic behavior of the right-hand side of the above equality. Note that the noncentral F-distribution can be written in terms of the ratio of two independent chi-square random

variables as

$$F(d, n-1-d; \rho_n^2) \overset{d}{=} \frac{\chi_d^2(\rho_n^2)/d}{\chi_{n-1-d}^2/(n-1-d)}.$$

For notational convenience, let us write $\mathcal{V}_{n,d} \overset{\text{def}}{=} \chi_{n-1-d}^2/(n-1-d)$. Then, by the weak law of large number, it is clear to see that $\mathcal{V}_{n,d} \xrightarrow{p} 1$ as $n, d \to \infty$ with $d/n \to 0$.

Main proof Our main strategy to prove the given claim is to split the cases into two: 1) $\rho_n^2/n \to 0$ and 2) $\liminf_{n,d\to\infty} \rho_n^2/n > 0$. In the first case, we shall show that $\chi_d^2(\rho_n^2)$ and $F(d, n-1-d; \rho_n^2)$ have the same asymptotic distribution after proper studentization. In the second case, we will verify that the power of both tests converge to one.

• **Case 1.** To begin, we assume $\rho_n^2/n \to 0$ and prove that

$$\frac{\chi_d^2(\rho_n^2)/\mathcal{V}_{n,d} - d - \rho_n^2}{\sqrt{2d+4\rho_n^2}} = \frac{\chi_d^2(\rho_n^2) - d - \rho_n^2}{\sqrt{2d+4\rho_n^2}} + o_P(1) \xrightarrow{d} N(0,1). \tag{F.5}$$

If (F.5) holds, then the result follows since

$$\frac{\frac{n_0+n_1-1-d}{n_0+n_1-2}\frac{n_0 n_1}{n} T_H - d - \rho_n^2}{\sqrt{2d+4\rho_n^2}} \xrightarrow{d} N(0,1) \quad \text{and} \quad \frac{dq_{\alpha,n,d} - d}{\sqrt{2d}} \to z_\alpha.$$

To show (F.5), note that a simple algebraic manipulation yields

$$\frac{\chi_d^2(\rho_n^2)/\mathcal{V}_{n,d} - d - \rho_n^2}{\sqrt{2d+4\rho_n^2}} = \frac{1}{\mathcal{V}_{n,d}} \left[\frac{\chi_d^2(\rho_n^2) - d - \rho_n^2}{\sqrt{2d+4\rho_n^2}}\right] + \frac{d+\rho_n^2}{\sqrt{2d+4\rho_n^2}} \left(\frac{1}{\mathcal{V}_{n,d}} - 1\right). \tag{F.6}$$

Using the moments of an inverse chi-square distribution,

$$\mathbb{E}\left[\frac{1}{\mathcal{V}_{n,d}}\right] = \frac{n-1-d}{n-3-d}, \quad \text{and} \quad \mathbb{V}\left[\frac{1}{\mathcal{V}_{n,d}}\right] = \frac{2(n-1-d)^2}{(n-3-d)^2(n-5-d)},$$

one can conclude that

$$\frac{d+\rho_n^2}{\sqrt{2d+4\rho_n^2}} \left(\frac{1}{\mathcal{V}_{n,d}} - 1\right) \xrightarrow{p} 0.$$

Then the result follows by Slutsky's theorem combined with $\mathcal{V}_{n,d} \xrightarrow{p} 1$ and (F.4).

• **Case 2.** In the second case where $\liminf_{n,d\to\infty} \rho_n^2/n > 0$, there is no guarantee of (F.5). Nevertheless, we can show that the power of both tests converge to one when $\liminf_{n,d\to\infty} \rho_n^2/n > 0$. Since the first term in

(7.7) is bounded and

$$\frac{\rho_n^2}{\sqrt{2d+4\rho_n^2}} \to \infty,$$

we can conclude that the power of φ_H^* converges to one when $\liminf_{n,d\to\infty} \rho_n^2/n > 0$.

Now we compute the limiting power of the test based on T_H. By putting

$$r_{n,d} = \frac{n_0+n_1-1-d}{n_0+n_1-2}\frac{n_0 n_1}{n_0+n_1},$$

one can note that

$$\begin{aligned}
&\mathbb{P}\left(\frac{r_{n,d}T_H - d}{\sqrt{2d}} \geq \frac{q_{\alpha,n,d}-d}{\sqrt{2d}}\right) \\
&= \mathbb{P}\left(\frac{r_{n,d}T_H - \mathbb{E}[r_{n,d}T_H]}{\sqrt{\mathbb{V}[r_{n,d}T_H]}} > \frac{q_{\alpha,n,d}-d}{\sqrt{2d}}\sqrt{\frac{2d}{\mathbb{V}[r_{n,d}T_H]}} + \frac{d-\mathbb{E}[r_{n,d}T_H]}{\sqrt{\mathbb{V}[r_{n,d}T_H]}}\right).
\end{aligned}$$

Using the mean and variance formula for a noncentral F-distribution, we have

$$\begin{aligned}
\mathbb{E}[r_{n,d}T_H] &= \frac{(n-1-d)(d+\rho_n^2)}{n-3-d} = (d+\rho_n^2)\{1+o(1)\}, \text{ and} \\
\mathbb{V}[r_{n,d}T_H] &= 2(n-1-d)^2\frac{(d+\rho_n^2)^2 + (d+2\rho_n^2)(n-3-d)}{(n-3-d)^2(n-5-d)} \\
&\asymp \frac{(d+\rho_n^2)^2}{n} + \frac{d+\rho_n^2}{n^2}.
\end{aligned}$$

From this, we may infer that

$$\begin{aligned}
&\frac{r_{n,d}T_H - \mathbb{E}[r_{n,d}T_H]}{\sqrt{\mathbb{V}[r_{n,d}T_H]}} = O_P(1), \quad \frac{q_{\alpha,n,d}-d}{\sqrt{2d}}\sqrt{\frac{2d}{\mathbb{V}[r_{n,d}T_H]}} = O(1) \text{ and} \\
&\frac{d-\mathbb{E}[r_{n,d}T_H]}{\sqrt{\mathbb{V}[r_{n,d}T_H]}} \to -\infty.
\end{aligned}$$

This immediately implies that

$$\liminf_{n,d\to\infty} \mathbb{P}\left(\frac{r_{n,d}T_H - d}{\sqrt{2d}} \geq \frac{q_{\alpha,n,d}-d}{\sqrt{2d}}\right) = 1,$$

thus completing the proof of Theorem 7.1.

F.3.4 Proof of Proposition 7.2 (asymptotic normality of W_A)

As described in the main text, the sample-splitting error of the generalized LDA classifier is an average of independent (but not all identically distributed) random variables when conditioning on the training set. Hence we apply the Berry-Esseen theorem in Lemma F.0.2 to first establish the conditional central limit theorem for W_A. Then we use the bounded convergence theorem to prove the unconditional counterpart.

Conditional Part Conditional on the training set $\mathcal{T}_{\text{tr}} \overset{\text{def}}{=} \mathcal{X}_1^{n_{0,\text{tr}}} \cup \mathcal{Y}_1^{n_{1,\text{tr}}}$, $\widehat{E}_A^S$ is the sum of independent random variables. Specifically,

$$
\begin{aligned}
n_{\text{te}}\widehat{E}_A^S &= \sum_{i=1}^{n_{0,\text{te}}} Q_{0,i} + \sum_{i=1}^{n_{1,\text{te}}} Q_{1,i}, \\
\text{where} \quad Q_{0,i} &= \frac{n_{\text{te}}}{2n_{0,\text{te}}} \mathbb{I}\Big[\text{LDA}_{A,n_{0,\text{tr}},n_{1,\text{tr}}}(X_{n_{0,\text{tr}}+i}) = 1\Big], \\
\text{and} \quad Q_{1,i} &= \frac{n_{\text{te}}}{2n_{1,\text{te}}} \mathbb{I}\Big[\text{LDA}_{A,n_{0,\text{tr}},n_{1,\text{tr}}}(Y_{n_{1,\text{tr}}+i}) = 0\Big].
\end{aligned}
$$

Notice that for $k = 0, 1$, we have

$$
\begin{aligned}
\mathbb{E}[|Q_{k,i} - \mathbb{E}[Q_{k,i}|\mathcal{T}_{\text{tr}}]|^3|\mathcal{T}_{\text{tr}}] &\leq \frac{n_{\text{te}}^3}{8n_{k,\text{te}}^3}, \text{ and} \\
\mathbb{E}[(Q_{k,i} - \mathbb{E}[Q_{k,i}|\mathcal{T}_{\text{tr}}])^2|\mathcal{T}_{\text{tr}}] &= \frac{n_{\text{te}}^2}{4n_{k,\text{te}}^2}\mathcal{E}_{1,A}(1-\mathcal{E}_{1,A}).
\end{aligned}
$$

We may then apply Lemma F.0.2 to yield

$$
\sup_{t\in\mathbb{R}} |\Pr(W_A \leq t|\mathcal{T}_{\text{tr}}) - \Phi(t)| \leq C_1 a_n b_n c_n, \tag{F.7}
$$

$$
\begin{aligned}
\text{where} \quad a_n &= \left\{\frac{n_{\text{te}}^2}{4n_{0,\text{te}}}\mathcal{E}_{0,A}(1-\mathcal{E}_{0,A}) + \frac{n_{\text{te}}^2}{4n_{1,\text{te}}}\mathcal{E}_{1,A}(1-\mathcal{E}_{1,A})\right\}^{-1/2}, \\
b_n &= \frac{n_{\text{te}}^3}{8n_{0,\text{te}}^3} + \frac{n_{\text{te}}^3}{8n_{1,\text{te}}^3}, \\
\text{and} \quad c_n &= \frac{4n_{0,\text{te}}^2}{n_{\text{te}}^2\mathcal{E}_{0,A}(1-\mathcal{E}_{0,A})} + \frac{4n_{1,\text{te}}^2}{n_{\text{te}}^2\mathcal{E}_{1,A}(1-\mathcal{E}_{1,A})}.
\end{aligned}
$$

Under the eigenvalue conditions for A and Σ in **(A5)** and **(A6)**, one can find constants $C_0, C_1 > 0$ such that $C_0 \leq \text{tr}\{(A\Sigma)^2\}/d \leq C_1$ and $C_0 \leq \text{tr}(A\Sigma)/d \leq C_1$ due to $d\lambda_{\min}^2(A)\lambda_{\min}^2(A) \leq \text{tr}\{(A\Sigma)^2\} \leq d\lambda_{\max}^2(A)\lambda_{\max}^2(A)$ and $d\lambda_{\min}(A)\lambda_{\min}(A) \leq \text{tr}(A\Sigma) \leq d\lambda_{\max}(A)\lambda_{\max}(A)$. Then under **(A1)**–**(A4)**, there exists another constant $C_2 > 0$ such that $-C_2 \leq (\Psi_{A,n,d} + \Xi_{A,n,d})/\sqrt{\Lambda_{A,n,d}} \leq C_2$ and $-C_2 \leq (\Psi_{A,n,d} - \Xi_{A,n,d})/\sqrt{\Lambda_{A,n,d}} \leq C_2$ for large n. Therefore both $\Phi\{(\Psi_{A,n,d}+\Xi_{A,n,d})/\sqrt{\Lambda_{A,n,d}}\}$ and $\Phi\{(\Psi_{A,n,d} -$

$\Xi_{A,n,d})/\sqrt{\Lambda_{A,n,d}}\}$ are strictly bounded below by zero and above by one for large n. Based on this observation together with Lemma F.0.4, it can be seen that $a_n = O_P(n^{-1/2})$, $b_n = O(1)$ and $c_n = O_P(1)$. Thus the right-hand side of (F.7) is $O_P(n^{-1/2})$, which completes the proof of the conditional part.

Unconditional Part For each $t \in \mathbb{R}$, the previous result gives $\Pr(W_A \leq t|\mathcal{X}_1^{n_{0,\mathrm{tr}}}, \mathcal{Y}_1^{n_{1,\mathrm{tr}}}) - \Phi(t) = o_P(1)$. We then apply the bounded convergence theorem to have $\Pr(W_A \leq t) - \Phi(t) = o(1)$. Since $\Phi(\cdot)$ is continuous, Polya's theorem yields the final result (e.g. Lemma 2.11 of Van der Vaart, 2000). This completes the proof of Proposition 7.2.

F.3.5 Proof of Theorem 7.2 and 7.4

Proof of Theorem 7.2 (Asymptotic normality of $W_A^\dagger$) Based on Lemma F.0.4 and the facts that $\Psi_{A,n,d} = O(n^{-1/2})$, $\Xi_{A,n,d} = O(1)$ and $\liminf_{n,d\to\infty} \Lambda_{A,n,d} > 0$ (see the proof of Proposition 7.2 for details), we have

$$\begin{aligned}\mathcal{E}_{0,A}(1-\mathcal{E}_{0,A}) &= \Phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\left\{1-\Phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\right\} + O_P(n^{-1/2}) \\ &= \Phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\left\{1-\Phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\right\} + O_P(n^{-1/2}), \end{aligned} \tag{F.8}$$

where the second line uses the first-order Taylor expansion:

$$\begin{aligned}\Phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) &= \Phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + O\left(\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) \\ &= \Phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + O(n^{-1/2}).\end{aligned}$$

Similarly, one can obtain

$$\mathcal{E}_{1,A}(1-\mathcal{E}_{1,A}) = \Phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\left\{1-\Phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\right\} + O_P(n^{-1/2}). \tag{F.9}$$

Then by substituting (F.8) and (F.9) into the definition of W_A,

$$\begin{aligned}W_A &= \frac{\widehat{E}_A^S - \mathcal{E}_{0,A}/2 - \mathcal{E}_{1,A}/2}{\sqrt{\mathcal{E}_{0,A}(1-\mathcal{E}_{0,A})/(4n_{0,\mathrm{te}}) + \mathcal{E}_{1,A}(1-\mathcal{E}_{1,A})/(4n_{1,\mathrm{te}})}} \\ &= 2\sqrt{\frac{n_{0,\mathrm{te}}n_{1,\mathrm{te}}}{n_{0,\mathrm{te}}+n_{1,\mathrm{te}}}}\times\end{aligned}$$

$$\frac{\widehat{E}_A^S - \mathcal{E}_{0,A}/2 - \mathcal{E}_{1,A}/2}{\sqrt{\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\{1-\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\} + O_P(n^{-1/2})}}. \tag{F.10}$$

By the Taylor expansion,

$$\begin{aligned}
&\frac{1}{\sqrt{\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\{1-\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\} + O_P(n^{-1/2})}} \\
= \ &\frac{1}{\sqrt{\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\{1-\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\}}} + O_P(n^{-1/2}).
\end{aligned}$$

Now by plugging this into (F.10) and using the fact that $\widehat{E}_A^S - \mathcal{E}_{0,A}/2 - \mathcal{E}_{1,A}/2 = O_P(n^{-1/2})$, one can obtain that

$$\begin{aligned}
W_A \ = \ & 2\sqrt{\frac{n_{0,\mathrm{te}} n_{1,\mathrm{te}}}{n_{0,\mathrm{te}} + n_{1,\mathrm{te}}}} \times \\
& \frac{\widehat{E}_A^S - \mathcal{E}_{0,A}/2 - \mathcal{E}_{1,A}/2}{\sqrt{\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\{1-\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\}}} + O_P(n^{-1/2}).
\end{aligned}$$

Lemma F.0.4 further allows us to replace $\mathcal{E}_{0,A}/2 + \mathcal{E}_{1,A}/2$ with its non-random counterpart as

$$\begin{aligned}
W_A = 2\sqrt{\frac{n_{0,\mathrm{te}} n_{1,\mathrm{te}}}{n_{0,\mathrm{te}} + n_{1,\mathrm{te}}}} & \frac{1}{\sqrt{\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\{1-\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\}}} \times \\
& \left\{ \widehat{E}_A^S - \frac{1}{2}\Phi\left(\frac{\Psi_{A,n,d} + \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + \frac{1}{2}\Phi\left(\frac{\Psi_{A,n,d} - \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) \right\} + o_P(1).
\end{aligned} \tag{F.11}$$

Additionally, using Taylor expansion of $\Phi(x)$ around $x = \Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}}$ or $x = -\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}}$, it is seen that

$$\begin{aligned}
\Phi\left(\frac{\Psi_{A,n,d} + \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) &= \Phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + \phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}} + o(n^{-1/2}), \\
\Phi\left(\frac{\Psi_{A,n,d} - \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) &= \Phi\left(\frac{-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + \phi\left(\frac{-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}} + o(n^{-1/2}).
\end{aligned}$$

This, together with $\Phi(x) + \Phi(-x) = 1$ and $\phi(x) = \phi(-x)$, gives

$$\begin{aligned}
&\Phi\left(\frac{\Psi_{A,n,d} + \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + \Phi\left(\frac{\Psi_{A,n,d} - \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) \\
= \ &1 + 2\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}} + o(n^{-1/2}).
\end{aligned} \tag{F.12}$$

Now combining (F.11) with (F.12), our final approximation is given by $W_A = W_A^{\dagger} + o_P(1)$. This proves the first part of Theorem 7.2.

For the second part, since $\Phi(x) + \Phi(-x) = 1$ for all $x \in \mathbb{R}$ and $\Psi_{A,n,d} = -\delta^\top A\delta/2 = o(1)$, we have that

$$\Phi\left(\frac{\Psi_{A,n,d} + \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + \Phi\left(\frac{\Psi_{A,n,d} - \Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) = 1 + o(1).$$

Thus the result follows by Lemma F.0.4, which completes the proof of Theorem 7.2.

Proof of Theorem 7.4 (Power of the naive Bayes classifier test) Based on Lemma F.0.5, one can establish as in Theorem 7.2 that

$$\sqrt{2n}\left(\widehat{E}^S_{\widehat{D}^{-1}} - \frac{1}{2} - \frac{\Psi_{D^{-1},n,d}}{\sqrt{2\pi\Lambda_{D^{-1},n,d}}}\right) \xrightarrow{d} N(0,1),$$

where we used $n_0 = n_1$, $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$ and $n_{\mathrm{tr}} = n_{\mathrm{te}}$. It is then straightforward to derive the power as in Section 7.6. Hence the result follows.

F.3.6 Proof of Lemma F.0.4

The proof of Lemma F.0.4 consists of three parts. In Part 1, we provide approximations of $V_{0,A}$, $V_{1,A}$ and $V_{0,A} + V_{1,A}$, which are defined around (7.11) (also recalled in Section F.3.1). In Part 2, we focus on U_A and present its approximation. In Part 3, by building on the results from the first two parts, we prove the main statements of Lemma F.0.4.

- **Part 1.**

Using Fan's inequality in Lemma F.0.3 under **(A5)** and **(A6)**, observe that

$$\mathrm{tr}\big\{(A\Sigma)^2\big\} = \mathrm{tr}(A\Sigma A\Sigma) \le d\lambda^2_{\max}(A)\lambda^2_{\max}(\Sigma) \lesssim d. \tag{F.13}$$

Then under **(A1)**, $\mathrm{tr}\big\{(A\Sigma)^2\big\} = O(n)$. Next based on the sub-multiplicative property of the operator norm and **(A2)**,

$$0 \le \delta^\top A\Sigma A\delta \le \lambda_{\max}(A\Sigma A)\delta^\top\delta \le \lambda^2_{\max}(A)\lambda_{\max}(\Sigma)\delta^\top\delta = O(n^{-1/2}). \tag{F.14}$$

Similarly, one can show that

$$\delta^\top A\Sigma A\Sigma A\Sigma A\delta = O(n^{-1/2}). \tag{F.15}$$

Using the ingredients above, we shall prove

$$\begin{aligned} V_{0,A} &= -\frac{1}{2}\delta^\top A\delta + \frac{1}{2}\left(\frac{1}{n_{0,\mathrm{tr}}} - \frac{1}{n_{1,\mathrm{tr}}}\right)\mathrm{tr}(A\Sigma) + O_P(n^{-1/2}), \\ V_{1,A} &= -\frac{1}{2}\delta^\top A\delta + \frac{1}{2}\left(\frac{1}{n_{1,\mathrm{tr}}} - \frac{1}{n_{0,\mathrm{tr}}}\right)\mathrm{tr}(A\Sigma) + O_P(n^{-1/2}), \quad \text{and} \\ V_{0,A} + V_{1,A} &= -\delta^\top A\delta + O_P(n^{-3/4}), \end{aligned} \tag{F.16}$$

where $V_{0,A}$ and $V_{1,A}$ are defined around (7.11). To this end, we need to calculate the mean and variance of $V_{0,A}$ and $V_{1,A}$. The calculation of the mean is rather straightforward as

$$\begin{aligned} \mathbb{E}[V_{0,A}] &= -\frac{1}{2}\delta^\top A\delta + \frac{1}{2}\left(\frac{1}{n_{0,\mathrm{tr}}} - \frac{1}{n_{1,\mathrm{tr}}}\right)\mathrm{tr}(A\Sigma), \\ \mathbb{E}[V_{1,A}] &= -\frac{1}{2}\delta^\top A\delta + \frac{1}{2}\left(\frac{1}{n_{1,\mathrm{tr}}} - \frac{1}{n_{0,\mathrm{tr}}}\right)\mathrm{tr}(A\Sigma). \end{aligned}$$

Turning to the variances, we will show that $\mathrm{Var}[V_{0,A}] = \mathrm{Var}[\widehat{\delta}^\top A(\mu_0 - \widehat{\mu}_{\mathrm{pool}})]$ is $O(n^{-1})$. First note that $(\widehat{\delta}, \mu_0 - \widehat{\mu}_{\mathrm{pool}})^\top$ has a multivariate normal distribution as

$$\begin{pmatrix} \widehat{\delta} \\ \mu_0 - \widehat{\mu}_{\mathrm{pool}} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 - \mu_0 \\ \frac{1}{2}\mu_0 - \frac{1}{2}\mu_1 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right),$$

where

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} (n_{0,\mathrm{tr}}^{-1} + n_{1,\mathrm{tr}}^{-1})\Sigma & \frac{1}{2}(n_{1,\mathrm{tr}}^{-1} - n_{0,\mathrm{tr}}^{-1})\Sigma \\ \frac{1}{2}(n_{1,\mathrm{tr}}^{-1} - n_{0,\mathrm{tr}}^{-1})\Sigma & \frac{1}{4}(n_{0,\mathrm{tr}}^{-1} + n_{1,\mathrm{tr}}^{-1})\Sigma \end{pmatrix}.$$

We also note that the conditional distribution of $\mu_0 - \widehat{\mu}_{\mathrm{pool}}$ given $\widehat{\delta}$ follows

$$\mu_0 - \widehat{\mu}_{\mathrm{pool}}|\widehat{\delta} \sim N\left(\mu^*, \Sigma^*\right), \tag{F.17}$$

where $\mu^* = -\delta/2 + \Sigma_{21}\Sigma_{11}^{-1}(\widehat{\delta} - \delta)$ and $\Sigma^* = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Next, by the law of total variance,

$$\mathrm{Var}[\widehat{\delta}^\top A(\mu_0 - \widehat{\mu}_{\mathrm{pool}})] = \underbrace{\mathbb{E}[\mathrm{Var}[\widehat{\delta}^\top A(\mu_0 - \widehat{\mu}_{\mathrm{pool}})|\widehat{\delta}]]}_{(I)} + \underbrace{\mathrm{Var}[\mathbb{E}[\widehat{\delta}^\top A(\mu_0 - \widehat{\mu}_{\mathrm{pool}})|\widehat{\delta}]]}_{(II)}.$$

Using (F.17), (I) and (II) are simplified as

$$(I) = \mathbb{E}[\widehat{\delta}^\top A\Sigma^* A\widehat{\delta}] \quad \text{and} \quad (II) = \mathrm{Var}[\widehat{\delta}^\top A\{-\delta/2 + \Sigma_{21}\Sigma_{11}^{-1}(\widehat{\delta} - \delta)\}].$$

By recalling the definitions of Σ^*, Σ_{21} and Σ_{11},

$$\begin{aligned}(I) \;\lesssim\; & \left(\frac{1}{n_{0,\mathrm{tr}}}+\frac{1}{n_{1,\mathrm{tr}}}\right)\mathbb{E}[\widehat{\delta}^\top A\Sigma A\widehat{\delta}] \\ & +\left(\frac{1}{n_{1,\mathrm{tr}}}-\frac{1}{n_{0,\mathrm{tr}}}\right)^2\left(\frac{1}{n_{0,\mathrm{tr}}}+\frac{1}{n_{1,\mathrm{tr}}}\right)^{-1}\mathbb{E}[\widehat{\delta}^\top A\Sigma A\widehat{\delta}] \\ \lesssim\; & \frac{1}{n}\mathbb{E}[\widehat{\delta}^\top A\Sigma A\widehat{\delta}], \end{aligned} \tag{F.18}$$

where the second line uses the assumptions **(A3)** and **(A4)**. Here the symbol $a_n \lesssim b_n$ means that there exists a constant $C>0$ such that $a_n \le Cb_n$ for large n. In addition, it can be checked that

$$(II) \;\lesssim\; \mathrm{Var}[\widehat{\delta}^\top A\delta]+\mathrm{Var}[\widehat{\delta}^\top A\widehat{\delta}]. \tag{F.19}$$

Now, based on Lemma F.0.1, one can verify that

$$\begin{aligned}\mathbb{E}[\widehat{\delta}^\top A\Sigma A\widehat{\delta}] &= \left(\frac{1}{n_{0,\mathrm{tr}}}+\frac{1}{n_{1,\mathrm{tr}}}\right)\mathrm{tr}(A\Sigma A\Sigma)+\delta^\top A\Sigma A\delta, \\ \mathrm{Var}[\widehat{\delta}^\top A\delta] &= \left(\frac{1}{n_{0,\mathrm{tr}}}+\frac{1}{n_{1,\mathrm{tr}}}\right)\delta^\top A\Sigma A\delta, \\ \mathrm{Var}[\widehat{\delta}^\top A\widehat{\delta}] &= 2\left(\frac{1}{n_{0,\mathrm{tr}}}+\frac{1}{n_{1,\mathrm{tr}}}\right)^2\mathrm{tr}(A\Sigma A\Sigma)+4\left(\frac{1}{n_{0,\mathrm{tr}}}+\frac{1}{n_{1,\mathrm{tr}}}\right)\delta^\top A\Sigma A\delta. \end{aligned}$$

By substituting the above expressions into (F.18) and (F.19) together with the preliminaries in (F.13) and (F.14), we have that $\mathrm{Var}[V_{0,A}] = O(n^{-1})$ as desired. The same lines of argument also show that $\mathrm{Var}[V_{1,A}] = O(n^{-1})$ and therefore the first two lines in (F.16) follow. Additionally, by noting that $V_{0,A}+V_{1,A} = \widehat{\delta}^\top A(\mu_0-\mu_1)$, we have

$$\begin{aligned}\mathbb{E}[V_{0,A}+V_{1,A}] &= -\delta^\top A\delta, \\ \mathrm{Var}[V_{0,A}+V_{1,A}] &= \left(\frac{1}{n_{0,\mathrm{tr}}}+\frac{1}{n_{1,\mathrm{tr}}}\right)\delta^\top A\Sigma A\delta. \end{aligned}$$

The above means and variances, together with (F.13) and (F.14), yield the claim (F.16).

- **Part 2.**

Applying Lemma F.0.1 yields

$$\begin{aligned}\mathbb{E}[U_A] &= \delta^\top A\Sigma A\delta + \left(\frac{1}{n_{0,\mathrm{tr}}} + \frac{1}{n_{1,\mathrm{tr}}}\right)\mathrm{tr}\big\{(A\Sigma)^2\big\} \quad \text{and} \\ \mathrm{Var}[U_A] &= 2\left(\frac{1}{n_{0,\mathrm{tr}}} + \frac{1}{n_{1,\mathrm{tr}}}\right)^2 \mathrm{tr}\big\{(A\Sigma)^4\big\} + 4\left(\frac{1}{n_{0,\mathrm{tr}}} + \frac{1}{n_{1,\mathrm{tr}}}\right)\delta^\top A\Sigma A\Sigma A\Sigma A\delta.\end{aligned} \tag{F.20}$$

As in (F.13), Fan's inequality shows $\mathrm{tr}\big\{(A\Sigma)^4\big\} = O(n)$. This fact, together with (F.15) and (F.20), gives

$$U_A = \delta^\top A\Sigma A\delta + \left(\frac{1}{n_{0,\mathrm{tr}}} + \frac{1}{n_{1,\mathrm{tr}}}\right)\mathrm{tr}\big\{(A\Sigma)^2\big\} + O_P(n^{-1/2}), \tag{F.21}$$

which completes the second part.

• **Part 3.**

Consider a bivariate function $f(v,u) = \Phi(v/\sqrt{u})$. Recall the definition of $\Psi_{A,n,d}$, $\Lambda_{A,n,d}$ and $\Xi_{A,n,d}$ in (7.12) (also recalled in Section F.3.1). Then by the Taylor expansion of $f(v,u)$ around $(\Psi_{A,n,d}+\Xi_{A,n,d}, \Lambda_{A,n,d})$ together with (F.16) and (F.21), we have

$$\begin{aligned}\mathcal{E}_{0,A} &= f(V_{0,A}, U_A) \\ &= \Phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + \phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{1}{\sqrt{\Lambda_{A,n,d}}}(V_{0,A} - \Psi_{A,n,d} - \Xi_{A,n,d}) \\ &\quad - \phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{(\Lambda_{A,n,d})^{3/2}}(U_A - \Lambda_{A,n,d}) + O_P\left(n^{-1}\right),\end{aligned} \tag{F.22}$$

where we recall that $\phi(\cdot)$ is the density function of $N(0,1)$. Similarly,

$$\begin{aligned}\mathcal{E}_{1,A} &= f(V_{1,A}, U_A) \\ &= \Phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + \phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{1}{\sqrt{\Lambda_{A,n,d}}}(V_{1,A} - \Psi_{A,n,d} + \Xi_{A,n,d}) \\ &\quad - \phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{(\Lambda_{A,n,d})^{3/2}}(U_A - \Lambda_{A,n,d}) + O_P\left(n^{-1}\right).\end{aligned} \tag{F.23}$$

Since the normal density function $\phi(\cdot)$ is bounded and $\liminf_{n,d\to\infty}\Lambda_{A,n,d}$ is a (strictly) positive constant under the given conditions (see the proof of Proposition 7.2), the first two claims in Lemma F.0.4 follow, i.e.

$$\mathcal{E}_{0,A} \;=\; \Phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + O_P\left(n^{-1/2}\right),$$

$$\mathcal{E}_{1,A} \;=\; \Phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)+O_P\left(n^{-1/2}\right).$$

Combining (F.22) and (F.23) yields that

$$\begin{aligned}\mathcal{E}_{0,A}+\mathcal{E}_{1,A} \;=\; & \Phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)+\Phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\\ & +(I)'-(II)'+O_P(n^{-1}),\end{aligned} \tag{F.24}$$

where

$$\begin{aligned}(I)' \;=\; & \phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{1}{\sqrt{\Lambda_{A,n,d}}}(V_{0,A}-\Psi_{A,n,d}-\Xi_{A,n,d})\\ & +\phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{1}{\sqrt{\Lambda_{A,n,d}}}(V_{1,A}-\Psi_{A,n,d}+\Xi_{A,n,d}) \quad \text{and}\\ (II)' \;=\; & \phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{(\Lambda_{A,n,d})^{3/2}}(U_A-\Lambda_{A,n,d})\\ & +\phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{(\Lambda_{A,n,d})^{3/2}}(U_A-\Lambda_{A,n,d}).\end{aligned}$$

Focusing on $(I)'$, we use the fact that $\phi(x)=\phi(-x)$ to obtain

$$\begin{aligned}(I)' \;=\; & \phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{1}{\sqrt{\Lambda_{A,n,d}}}(V_{0,A}+V_{1,A}-2\Psi_{A,n,d})\\ & +\left[\phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)-\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\right]\frac{1}{\sqrt{\Lambda_{A,n,d}}}(V_{0,A}-\Psi_{A,n,d}-\Xi_{A,n,d})\\ & +\left[\phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)-\phi\left(\frac{-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\right]\frac{1}{\sqrt{\Lambda_{A,n,d}}}(V_{1,A}-\Psi_{A,n,d}+\Xi_{A,n,d}).\end{aligned}$$

By using the asymptotic relationships in (F.16) and $\Psi_{A,n,d}=o(1)$,

$$\begin{aligned}(I)' \;&=\; O(1)\cdot O_P(n^{-3/4})+o(1)\cdot O_P(n^{-1/2})+o(1)\cdot O_P(n^{-1/2})\\ &=\; o_P(n^{-1/2}).\end{aligned}$$

Similarly, one can establish by using (F.21) and $\Psi_{A,n,d} = o(1)$ that

$$\begin{aligned}(II)' &= \left[\phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{(\Lambda_{A,n,d})^{3/2}}\right.\\ &\qquad\left.+\,\phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{(\Lambda_{A,n,d})^{3/2}}\right](U_A-\Lambda_{A,n,d})\\ &= o(1)\cdot O_P(n^{-1/2}) = o_P(n^{-1/2}).\end{aligned}$$

Now by substituting these results to (F.24),

$$\mathcal{E}_{A,0}+\mathcal{E}_{A,1} = \Phi\left(\frac{\Psi_{A,n,d}+\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)+\Phi\left(\frac{\Psi_{A,n,d}-\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)+o_P(n^{-1/2}),$$

as desired. If $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$, then the approximations of $(I)'$ and $(II)'$ become much more straightforward with $\Xi_{A,n,d} = 0$. Indeed, one can infer that $(I)' = O_P(n^{-3/4})$ and $(II)' = O_P(n^{-3/4})$, which yields

$$\mathcal{E}_{A,0}+\mathcal{E}_{A,1} = 2\Phi\left(\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)+O_P(n^{-3/4}).$$

This completes the proof of Lemma F.0.4.

F.3.7 Some moments of (scaled) inverse chi-square random variables

In this section, we provide two lemmas (Lemma F.0.6 and Lemma F.0.7) where we present some moments of (scaled) inverse chi-square random variables. These results will be used to prove Lemma F.0.5. Throughout this section, we assume that $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$. Let us denote the diagonal elements of $\widehat{D}$ by $s_1^2,\ldots,s_d^2$ where

$$s_k^2 = \frac{1}{2(n_{0,\mathrm{tr}}-1)}\sum_{i=1}^{n_{0,\mathrm{tr}}}(X_{ik}-\overline{X}_k)^2+\frac{1}{2(n_{1,\mathrm{tr}}-1)}\sum_{i=1}^{n_{1,\mathrm{tr}}}(Y_{ik}-\overline{Y}_k)^2,$$

for $k = 1,\ldots,d$. Here $\overline{X}_k$ and $\overline{Y}_k$ are the sample means based on the training set, i.e. $\overline{X}_k = n_{0,\mathrm{tr}}^{-1}\sum_{i=1}^{n_{0,\mathrm{tr}}} X_{ik}$ and $\overline{Y}_k = n_{1,\mathrm{tr}}^{-1}\sum_{i=1}^{n_{1,\mathrm{tr}}} Y_{ik}$. Then by putting $\sigma_k^2 = [\Sigma]_{k,k}$, we have

$$\frac{1}{s_k^2} \sim \frac{n_{\mathrm{tr}}-2}{\sigma_k^2}\text{inv-}\chi^2_{n_{\mathrm{tr}}-2}, \text{ and } \frac{(n_{\mathrm{tr}}-2)s_k^2}{\sigma_k^2} \sim \chi^2_{n_{\mathrm{tr}}-2}, \tag{F.25}$$

where inv-$\chi^2_{n_{\mathrm{tr}}-2}$ represents the inverse-chi-squared distribution with $n_{\mathrm{tr}}-2$ degrees of freedom.
Let us investigate some moments of s_k^{-2}, which will be used to control the inverse of $\widehat{D}$.

Lemma F.0.6. *Suppose that $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$. Then under **(A4)**, some of non-central moments of s_k^{-2} are given by*

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{s_k^2}\right] &= \frac{n_{\mathrm{tr}}-2}{\sigma_k^2}\frac{1}{n_{\mathrm{tr}}-4} = \frac{1}{\sigma_k^2}\{1+O(n^{-1})\},\\
\mathbb{E}\left[\frac{1}{s_k^4}\right] &= \frac{(n_{\mathrm{tr}}-2)^2}{\sigma_k^4}\frac{1}{(n_{\mathrm{tr}}-4)(n_{\mathrm{tr}}-8)} = \frac{1}{\sigma_k^4}\{1+O(n^{-1})\},\\
\mathbb{E}\left[\frac{1}{s_k^6}\right] &= \frac{(n_{\mathrm{tr}}-2)^3}{\sigma_k^6}\frac{1}{(n_{\mathrm{tr}}-12)(n_{\mathrm{tr}}-8)(n_{\mathrm{tr}}-2)}\\
&= \frac{1}{\sigma_k^6}\{1+O(n^{-1})\},\\
\mathbb{E}\left[\frac{1}{s_k^8}\right] &= \frac{(n_{\mathrm{tr}}-8)^4}{\sigma_k^8}\frac{1}{(n_{\mathrm{tr}}-16)(n_{\mathrm{tr}}-12)(n_{\mathrm{tr}}-8)(n_{\mathrm{tr}}-4)}\\
&= \frac{1}{\sigma_k^8}\{1+O(n^{-1})\}.
\end{aligned}$$

In addition, a couple of the central moments are

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{s_k^2}-\mathbb{E}\left[\frac{1}{s_k^2}\right]\right)^2\right] &= \frac{(n_{\mathrm{tr}}-2)^2}{\sigma_k^4}\frac{2}{(n_{\mathrm{tr}}-4)^2(n_{\mathrm{tr}}-6)}\\
&= \frac{1}{\sigma_k^2}\cdot O(n^{-1}),\\
\mathbb{E}\left[\left(\frac{1}{s_k^2}-\mathbb{E}\left[\frac{1}{s_k^2}\right]\right)^4\right] &= \frac{(n_{\mathrm{tr}}-2)^4}{\sigma_k^8}\frac{12(n_{\mathrm{tr}}-2)^2+72(n_{\mathrm{tr}}-2)-480}{(n_{\mathrm{tr}}-8)(n_{\mathrm{tr}}-10)(n_{\mathrm{tr}}-4)^4(n_{\mathrm{tr}}-6)^2}\\
&= \frac{1}{\sigma_k^8}\cdot O(n^{-2}).
\end{aligned}$$

Proof. Suppose that $X \sim \chi^2_\nu$. Then for $\nu \geq 2k+1$,

$$\begin{aligned}
\mathbb{E}[X^{-k}] &= \int_0^\infty x^{-k}\frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{\nu/2-1}e^{-x/2}dx\\
&= \frac{1}{2^{2k}}\frac{\Gamma(\nu/2-k)}{\Gamma(\nu/2)}\int_0^\infty \frac{1}{2^{\frac{\nu-2k}{2}}\Gamma\{(\nu-2k)/2\}}x^{\frac{\nu-2k}{2}-1}e^{-x/2}dx\\
&= \frac{1}{2^{2k}}\frac{\Gamma(\nu/2-k)}{\Gamma(\nu/2)},
\end{aligned}$$

where the last equality uses the fact that a density integrates to one. Using this exact inverse moment expression and the relationship in (F.25), the results follows by straightforward algebra. □

Next we study some product moments of (scaled) inverse chi-square random variables.

Lemma F.0.7. *Suppose that $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$. Then under **(A4)**, for any $1 \le i, j, k, l \le d$,*

$$\mathbb{E}\left[\frac{1}{s_i^2 s_j^2} - \frac{1}{\sigma_i^2 \sigma_j^2}\right] = O(n^{-1}), \tag{F.26}$$

$$\mathbb{E}\left[\frac{1}{s_i^2 s_j^2 s_k^2} - \frac{1}{\sigma_i^2 \sigma_j^2 \sigma_k^2}\right] = O(n^{-1}) \text{ and} \tag{F.27}$$

$$\mathbb{E}\left[\frac{1}{s_i^2 s_j^2 s_k^2 s_l^2} - \frac{1}{\sigma_i^2 \sigma_j^2 \sigma_k^2 \sigma_l^2}\right] = O(n^{-1}). \tag{F.28}$$

Proof. To prove claim (F.26), write

$$\frac{1}{s_i^2 s_j^2} - \frac{1}{\sigma_i^2 \sigma_j^2} = \left(\frac{1}{s_i^2} - \frac{1}{\sigma_i^2}\right)\left(\frac{1}{s_j^2} - \frac{1}{\sigma_j^2}\right) + \left(\frac{1}{s_i^2} - \frac{1}{\sigma_i^2}\right)\frac{1}{\sigma_j^2} + \left(\frac{1}{s_j^2} - \frac{1}{\sigma_j^2}\right)\frac{1}{\sigma_i^2}.$$

Then by using Cauchy-Schwarz inequality, we see that

$$\begin{aligned}\left|\mathbb{E}\left[\frac{1}{s_i^2 s_j^2} - \frac{1}{\sigma_i^2 \sigma_j^2}\right]\right| \;\le\; & \mathbb{E}\left[\left(\frac{1}{s_i^2} - \frac{1}{\sigma_i^2}\right)^2\right] + \frac{1}{\sigma_j^2}\left|\mathbb{E}\left[\frac{1}{s_i^2} - \frac{1}{\sigma_i^2}\right]\right| \\ & + \frac{1}{\sigma_i^2}\left|\mathbb{E}\left[\frac{1}{s_j^2} - \frac{1}{\sigma_j^2}\right]\right|.\end{aligned} \tag{F.29}$$

The three terms on the right-hand side are $O(n^{-1})$ by Lemma F.0.6 and thus (F.26) follows.

Next we prove (F.27); the result in (F.28) follows similarly. Write

$$\begin{aligned}\frac{1}{s_i^2 s_j^2 s_k^2} - \frac{1}{\sigma_i^2 \sigma_j^2 \sigma_k^2} \;=\; & \left(\frac{1}{s_i^2 s_j^2} - \frac{1}{\sigma_i^2 \sigma_j^2}\right)\left(\frac{1}{s_k^2} - \frac{1}{\sigma_k^2}\right) \\ & + \left(\frac{1}{s_i^2 s_j^2} - \frac{1}{\sigma_i^2 \sigma_j^2}\right)\frac{1}{\sigma_k^2} + \left(\frac{1}{s_k^2} - \frac{1}{\sigma_k^2}\right)\frac{1}{\sigma_i^2 \sigma_j^2}.\end{aligned} \tag{F.30}$$

Note that the expected values of the last two terms in (F.30) are $O(n^{-1})$ by Lemma F.0.6 and (F.26). Therefore we focus on the first term and show that its expected value is $O(n^{-1})$. The first term can be decomposed as

$$\begin{aligned}& \left(\frac{1}{s_i^2 s_j^2} - \frac{1}{\sigma_i^2 \sigma_j^2}\right)\left(\frac{1}{s_k^2} - \frac{1}{\sigma_k^2}\right) \\ = \;& \left[\left(\frac{1}{s_i^2} - \frac{1}{\sigma_i^2}\right)\left(\frac{1}{s_j^2} - \frac{1}{\sigma_j^2}\right) + \left(\frac{1}{s_i^2} - \frac{1}{\sigma_i^2}\right)\frac{1}{\sigma_j^2} + \left(\frac{1}{s_j^2} - \frac{1}{\sigma_j^2}\right)\frac{1}{\sigma_i^2}\right]\left(\frac{1}{s_k^2} - \frac{1}{\sigma_k^2}\right).\end{aligned} \tag{F.31}$$

We only need to handle the following term in (F.31)

$$\left(\frac{1}{s_i^2}-\frac{1}{\sigma_i^2}\right)\left(\frac{1}{s_j^2}-\frac{1}{\sigma_j^2}\right)\left(\frac{1}{s_k^2}-\frac{1}{\sigma_k^2}\right), \tag{F.32}$$

since the expected values of the other terms are $O(n^{-1})$, which follows as in (F.29) using Cauchy-Schwarz inequality. But the expectation of (F.32) is $O(n^{-1})$ again by Cauchy-Schwarz inequality and Lemma F.0.6. Thus the expectation of (F.31) is $O(n^{-1})$. Finally, after substituting this result into the expectation of (F.30), we may obtain the result in (F.27). Hence Lemma F.0.7 follows. □

F.3.8 Proof of Lemma F.0.5

Let us denote

$$\begin{aligned} V_{0,\widehat{D}^{-1}} &\stackrel{\text{def}}{=} \widehat{\delta}^\top \widehat{D}^{-1}(\mu_0-\widehat{\mu}_{\text{pool}}), \\ V_{1,\widehat{D}^{-1}} &\stackrel{\text{def}}{=} \widehat{\delta}^\top \widehat{D}^{-1}(\widehat{\mu}_{\text{pool}}-\mu_1), \text{ and} \\ U_{\widehat{D}^{-1}} &\stackrel{\text{def}}{=} \widehat{\delta}^\top \widehat{D}^{-1}\Sigma\widehat{D}^{-1}\widehat{\delta}. \end{aligned}$$

By assuming **(A1)**–**(A5)** with $n_0 = n_1$, $n_{0,\text{tr}} = n_{1,\text{tr}}$ and $n_{\text{tr}} = n_{\text{te}}$, we break the proof up into three parts:

- **Part 1.** $V_{0,\widehat{D}^{-1}} = \Psi_{D^{-1},n,d} + O_P\left(n^{-1/2}\right)$ and $V_{1,\widehat{D}^{-1}} = \Psi_{D^{-1},n,d} + O_P\left(n^{-1/2}\right)$.

- **Part 2.** $U_{\widehat{D}^{-1}} = \Lambda_{D^{-1},n,d} + O_P\left(n^{-1/2}\right)$.

- **Part 3.** $V_{0,\widehat{D}^{-1}} + V_{1,\widehat{D}^{-1}} = 2\Psi_{D^{-1},n,d} + O_P\left(n^{-3/4}\right)$.

Suppose that the above claims hold. Then the final result of Lemma F.0.5 follows similarly as in the proof of Lemma F.0.4 via Taylor expansion. We now verify each claim in order.

- **Part 1.**

We only prove that

$$V_{0,\widehat{D}^{-1}} = -\frac{1}{2}\delta^\top D^{-1}\delta + O_P\left(\frac{1}{\sqrt{n}}\right).$$

The argument for $V_{1,\widehat{D}^{-1}}$ follows analogously. Under the Gaussian assumption with mutually independent random samples, one can see that the following vector

$$(\underbrace{\overline{X}_1, \ldots, \overline{X}_d, \overline{Y}_1, \ldots, \overline{Y}_d}_{(A)}, \underbrace{X_{11} - \overline{X}_1, \ldots, X_{n_0 1} - \overline{X}_1, Y_{11} - \overline{Y}_1, \ldots, Y_{n_1 d} - \overline{Y}_d}_{(B)})$$

has a multivariate normal distribution. Furthermore, a little algebra shows that the cross-covariance matrix between (A) and (B) is a zero matrix, which implies that (A) and (B) are independent under the Gaussian assumption. Since $\widehat{D}^{-1}$ is a function of (B), it shows that $\widehat{D}^{-1}$ is independent of $\widehat{\delta}$ and $\widehat{\mu}_{\text{pool}}$, which are functions of (A). In addition, when $n_{0,\text{tr}} = n_{1,\text{tr}}$, the covariance between $\widehat{\delta}$ and $\widehat{\mu}_{\text{pool}}$ is a zero matrix as

$$\text{Cov}(\widehat{\mu}_1 - \widehat{\mu}_0, \widehat{\mu}_1/2 + \widehat{\mu}_0/2) = \left(\frac{1}{2n_{1,\text{tr}}} - \frac{1}{2n_{0,\text{tr}}}\right)\Sigma = 0,$$

which further implies that $\widehat{D}^{-1}$, $\widehat{\delta}$ and $\widehat{\mu}_{\text{pool}}$ are mutually independent. Based on this observation, the expectation becomes

$$\begin{aligned}\mathbb{E}[V_{0,\widehat{D}^{-1}}] &= -\frac{1}{2}\delta^\top \mathbb{E}[\widehat{D}^{-1}]\delta = -\frac{1}{2}\sum_{i=1}^{d}\delta_i^2 \mathbb{E}\left[\frac{1}{s_i^2}\right] = -\frac{1}{2}\delta^\top D^{-1}\delta + \delta^\top\delta \cdot O(n^{-1}) \\ &= -\frac{1}{2}\delta^\top D^{-1}\delta + O\left(\frac{1}{n^{3/2}}\right), \end{aligned} \tag{F.33}$$

since $\delta^\top\delta = O(n^{-1/2})$ under **(A1)**, **(A2)** and **(A5)**.

Next calculate the second moment using Lemma F.0.1 as

$$\begin{aligned} &\mathbb{E}[V_{0,\widehat{D}^{-1}}^2] \\ =\ & \mathbb{E}\left[\text{tr}\left\{\mathbb{E}\left[\widehat{\delta}\widehat{\delta}^\top\right]\widehat{D}^{-1}\mathbb{E}\left[(\mu_0 - \widehat{\mu}_{\text{pool}})(\mu_0 - \widehat{\mu}_{\text{pool}})^\top\right]\widehat{D}^{-1}\right\}\right] \\ =\ & \mathbb{E}\left[\text{tr}\left\{\left[\delta\delta^\top + \frac{4}{n_{\text{tr}}}\Sigma\right]\widehat{D}^{-1}\left[\frac{1}{4}\delta\delta^\top + \frac{1}{n_{\text{tr}}}\Sigma\right]\widehat{D}^{-1}\right\}\right] \\ =\ & \underbrace{\frac{1}{4}\mathbb{E}\left[\left(\delta^\top\widehat{D}^{-1}\delta\right)^2\right]}_{(I)} + \underbrace{\frac{2}{n_{\text{tr}}}\mathbb{E}\left[\delta^\top\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\right]}_{(II)} + \underbrace{\frac{4}{n_{\text{tr}}^2}\mathbb{E}\left[\text{tr}\left\{\left(\Sigma\widehat{D}^{-1}\right)^2\right\}\right]}_{(III)}. \end{aligned}$$

For (I), we apply Lemma F.0.7 to have

$$(I) \ = \ \frac{1}{4}\sum_{i=1}^{d}\sum_{j=1}^{d}\delta_i^2\delta_j^2\mathbb{E}\left[\frac{1}{s_i^2 s_j^2}\right] = \frac{1}{4}\left(\delta^\top D^{-1}\delta\right)^2 + O\left(\frac{1}{n^2}\right).$$

For (II), by writing $\sigma_{ij} = [\Sigma]_{ij}$, we infer that

$$\begin{aligned}(II) &= \frac{2}{n_{\mathrm{tr}}}\mathbb{E}\left[\delta^\top \widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\right] = \frac{2}{n_{\mathrm{tr}}}\sum_{i=1}^{d}\sum_{j=1}^{d}\delta_i\delta_j\sigma_{ij}\mathbb{E}\left[\frac{1}{s_i^2 s_j^2}\right] \\ &= \frac{2}{n_{\mathrm{tr}}}\delta^\top D^{-1}\Sigma D^{-1}\delta + O\left(\frac{\delta^\top\Sigma\delta}{n^2}\right) = \frac{2}{n_{\mathrm{tr}}}\delta^\top D^{-1}\Sigma D^{-1}\delta + O\left(\frac{1}{n^{5/2}}\right).\end{aligned}$$

The last term simplifies as

$$\begin{aligned}(III) &= \frac{4}{n_{\mathrm{tr}}^2}\mathbb{E}\left[\mathrm{tr}\left\{\left(\Sigma\widehat{D}^{-1}\right)^2\right\}\right] = \frac{4}{n_{\mathrm{tr}}^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\sigma_{ij}^2\mathbb{E}\left[\frac{1}{s_i^2 s_j^2}\right] \\ &= \frac{4}{n_{\mathrm{tr}}^2}\mathrm{tr}\left\{\left(\Sigma D^{-1}\right)^2\right\} + O\left(\frac{\mathrm{tr}(\Sigma^2)}{n^3}\right) = \frac{4}{n_{\mathrm{tr}}^2}\mathrm{tr}\left\{\left(\Sigma D^{-1}\right)^2\right\} + O\left(\frac{1}{n^2}\right).\end{aligned}$$

Under the given assumptions, one can check that $\delta^\top D^{-1}\Sigma D^{-1}\delta = O(n^{-1/2})$ and $\mathrm{tr}\{(\Sigma D^{-1})^2\} = O(d)$. Thus

$$\mathbb{E}[V_{0,\widehat{D}^{-1}}^2] = (I) + (II) + (III) = \frac{1}{4}\left(\delta^\top D^{-1}\delta\right)^2 + O(n^{-1}),$$

which yields together with (F.33) that $\mathrm{Var}[V_{0,\widehat{D}^{-1}}] = O(n^{-1})$. Hence the result follows.

• **Part 2.**

First calculate the expectation. Conditioned on $\widehat{D}^{-1}$, apply Lemma F.0.1 to have

$$\begin{aligned}\mathbb{E}[U_{\widehat{D}^{-1}}] &= \mathbb{E}\Big[\mathbb{E}\Big[\widehat{\delta}^\top\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\widehat{\delta}|\widehat{D}\Big]\Big] \\ &= \underbrace{\mathbb{E}\Big[\delta^\top\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\Big]}_{(I)} + \underbrace{\frac{4}{n_{\mathrm{tr}}}\mathbb{E}\Big[\mathrm{tr}\left(\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\Sigma\right)\Big]}_{(II)}.\end{aligned}$$

For (I), by putting $\sigma_{ij} = [\Sigma]_{ij}$, we apply Lemma F.0.7 to have

$$(I) = \sum_{i=1}^{d}\sum_{j=1}^{d}\delta_i\delta_j\sigma_{ij}\mathbb{E}\left[\frac{1}{s_i^2 s_j^2}\right] = \delta^\top D^{-1}\Sigma D^{-1}\delta + O\left(\frac{1}{n^{3/2}}\right).$$

For (II),

$$(II) = \frac{4}{n_{\mathrm{tr}}}\sum_{i=1}^{d}\sum_{j=1}^{d}\sigma_{ij}^2\mathbb{E}\left[\frac{1}{s_i^2 s_j^2}\right] = \frac{4}{n_{\mathrm{tr}}}\mathrm{tr}\left(D^{-1}\Sigma D^{-1}\Sigma\right) + O\left(\frac{1}{n}\right).$$

Therefore

$$\mathbb{E}[U_{\widehat{D}^{-1}}] = \Lambda_{D^{-1},n,d} + O(n^{-1}).$$

Next compute the variance of $U_{\widehat{D}^{-1}}$.

$$\mathrm{Var}[U_{\widehat{D}^{-1}}] = \mathbb{E}\{\mathrm{Var}[U_{\widehat{D}^{-1}}|\widehat{D}]\} + \mathrm{Var}\{\mathbb{E}[U_{\widehat{D}^{-1}}|\widehat{D}]\}. \tag{F.34}$$

Using Lemma F.0.1 and the fact that $\widehat{\delta}$, $\widehat{D}^{-1}$ and $\widehat{\mu}_{\text{pool}}$ are mutually independent (see Part 1),

$$\mathrm{Var}[U_{\widehat{D}^{-1}}|\widehat{D}] = \frac{32}{n_{\text{tr}}^2}\mathrm{tr}\Big\{\Big(\widehat{D}^{-1}\Sigma\Big)^4\Big\} + \frac{16}{n_{\text{tr}}}\delta^\top\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta. \tag{F.35}$$

For the first term, we use Lemma F.0.7 to obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{32}{n_{\text{tr}}^2}\mathrm{tr}\Big\{\Big(\widehat{D}^{-1}\Sigma\Big)^4\Big\}\right] &= \frac{32}{n_{\text{tr}}^2}\mathbb{E}\left[\sum_{i=1}^d\sum_{j=1}^d\left(\sum_{k=1}^d\frac{\sigma_{ik}\sigma_{kj}}{s_i^2s_k^2}\right)^2\right] \\
&= \frac{32}{n_{\text{tr}}^2}\mathrm{tr}\Big\{\Big(D^{-1}\Sigma\Big)^4\Big\} + O\left(\frac{\mathrm{tr}\{\Sigma^4\}}{n^2}\right) \\
&= \frac{32}{n_{\text{tr}}^2}\mathrm{tr}\Big\{\Big(D^{-1}\Sigma\Big)^4\Big\} + O\left(\frac{1}{n}\right) \\
&= O\left(\frac{1}{n}\right).
\end{aligned} \tag{F.36}$$

Similarly for the second term,

$$\begin{aligned}
\frac{16}{n_{\text{tr}}}\mathbb{E}\Big[\delta^\top\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\Big] &= \frac{16}{n_{\text{tr}}}\delta^\top D^{-1}\Sigma D^{-1}\Sigma D^{-1}\Sigma D^{-1}\delta + O\left(\frac{\delta^\top\Sigma^3\delta}{n^2}\right) \\
&= O\left(\frac{1}{n^{3/2}}\right).
\end{aligned} \tag{F.37}$$

By substituting (F.36) and (F.37) into the the expectation of (F.35), we can conclude that

$$\mathbb{E}\{\mathrm{Var}[U_{\widehat{D}^{-1}}|\widehat{D}]\} = O(n^{-1}).$$

Returning to decomposition (F.34) and next focusing on $\mathrm{Var}\{\mathbb{E}[U_{\widehat{D}^{-1}}|\widehat{D}]\}$, note that

$$\mathbb{E}[U_{\widehat{D}^{-1}}|\widehat{D}] = \delta^\top\widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta + \frac{4}{n_{\text{tr}}}\mathrm{tr}\{(\widehat{D}^{-1}\Sigma)^2\}.$$

Thus

$$\mathrm{Var}\{\mathbb{E}[U_{\widehat{D}^{-1}}|\widehat{D}]\} \;\leq\; 2\underbrace{\mathrm{Var}\Big[\delta^\top \widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\Big]}_{(I)'} + 4\underbrace{\mathrm{Var}\Big[2n_{\mathrm{tr}}^{-1}\mathrm{tr}\big\{(\widehat{D}^{-1}\Sigma)^2\big\}\Big]}_{(II)'}. \tag{F.38}$$

For $(I)'$, we use Lemma F.0.7 to obtain

$$\begin{aligned}
\mathbb{E}\big[\big(\delta^\top \widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\big)^2\big] &= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{i'=1}^{d}\sum_{j'=1}^{d}\delta_i\delta_j\delta_{i'}\delta_{j'}\sigma_{ij}\sigma_{i'j'}\mathbb{E}\left[\frac{1}{s_i^2 s_j^2 s_{i'}^2 s_{j'}^2}\right] \\
&= \big(\delta^\top D^{-1}\Sigma D^{-1}\delta\big)^2 + O\left(\frac{\big(\delta^\top\Sigma\delta\big)^2}{n}\right) \\
&= \big(\delta^\top D^{-1}\Sigma D^{-1}\delta\big)^2 + O\left(\frac{1}{n^2}\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\big[\delta^\top \widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\big] &= \sum_{i=1}^{d}\sum_{j=1}^{d}\delta_i\delta_j\sigma_{ij}\mathbb{E}\left[\frac{1}{s_i^2 s_j^2}\right] \\
&= \delta^\top D^{-1}\Sigma D^{-1}\delta + O\left(\frac{\delta^\top\Sigma\delta}{n}\right) \\
&= \delta^\top D^{-1}\Sigma D^{-1}\delta + O\left(\frac{1}{n^{3/2}}\right).
\end{aligned}$$

Therefore, $(I)' = \mathbb{E}\big[\big(\delta^\top \widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\big)^2\big] - \big\{\mathbb{E}\big[\delta^\top \widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\big]\big\}^2 = O(n^{-2})$.

Moving onto $(II)'$, we have

$$\begin{aligned}
\mathbb{E}\big[\big(2n_{\mathrm{tr}}^{-1}\mathrm{tr}\big\{(\widehat{D}^{-1}\Sigma)^2\big\}\big)^2\big] &= 4n_{\mathrm{tr}}^{-2}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{i'=1}^{d}\sum_{j'=1}^{d}\sigma_{ij}^2\sigma_{i'j'}^2\mathbb{E}\left[\frac{1}{s_i^2 s_j^2 s_{i'}^2 s_{j'}^2}\right] \\
&= 4n_{\mathrm{tr}}^{-2}\big[\mathrm{tr}\big\{(D^{-1}\Sigma)^2\big\}\big]^2 + O\left(\frac{\big\{\mathrm{tr}(\Sigma^2)\big\}^2}{n^3}\right) \\
&= 4n_{\mathrm{tr}}^{-2}\big[\mathrm{tr}\big\{(D^{-1}\Sigma)^2\big\}\big]^2 + O(n^{-1}),
\end{aligned}$$

and

$$\mathbb{E}\big[2n_{\mathrm{tr}}^{-1}\mathrm{tr}\big\{(\widehat{D}^{-1}\Sigma)^2\big\}\big] \;=\; 2n_{\mathrm{tr}}^{-1}\sum_{i=1}^{d}\sum_{j=1}^{d}\sigma_{ij}^2\mathbb{E}\left[\frac{1}{s_i^2 s_j^2}\right]$$

$$= 2n_{\text{tr}}^{-1}\text{tr}\big\{(D^{-1}\Sigma)^2\big\} + O\left(\frac{\text{tr}\{\Sigma^2\}}{n^2}\right)$$

$$= 2n_{\text{tr}}^{-1}\text{tr}\big\{(D^{-1}\Sigma)^2\big\} + O(n^{-1}).$$

Hence $(II)' = O(n^{-1})$. Substituting the bounds $(I)' = O(n^{-2})$ and $(II)' = O(n^{-1})$ into the right-hand side of (F.38), we obtain that

$$\text{Var}[U_{\widehat{D}^{-1}}] = O(n^{-1}).$$

This verifies the second part.

• **Part 3.**

Let us start with the expectation. Based on the fact that $\widehat{\delta}$, $\widehat{D}^{-1}$ and $\widehat{\mu}_{\text{pool}}$ are mutually independent (see Part 1),

$$\begin{aligned}\mathbb{E}[V_{0,\widehat{D}^{-1}} + V_{1,\widehat{D}^{-1}}] &= \mathbb{E}\left[\widehat{\delta}^\top \widehat{D}^{-1}(\mu_0 - \widehat{\mu}_{\text{pool}}) + \widehat{\delta}^\top \widehat{D}^{-1}(\widehat{\mu}_{\text{pool}} - \mu_1)\right] \\ &= -\delta^\top D^{-1}\delta \cdot \big\{1 + O(n^{-1})\big\}.\end{aligned}$$

Next calculate the second moment.

$$\begin{aligned}\mathbb{E}[(V_{0,\widehat{D}^{-1}} + V_{1,\widehat{D}^{-1}})^2] &= \mathbb{E}\left[\text{tr}\left(\widehat{\delta}\widehat{\delta}^\top \widehat{D}^{-1}\delta\delta^\top \widehat{D}^{-1}\right)\right] \\ &= \mathbb{E}\left[\text{tr}\Big\{\left(\delta\delta^\top + 4n_{\text{tr}}^{-1}\Sigma\right)\widehat{D}^{-1}\delta\delta^\top \widehat{D}^{-1}\Big\}\right] \\ &= \underbrace{\mathbb{E}\left[\left(\delta^\top \widehat{D}^{-1}\delta\right)^2\right]}_{(I)''} + \underbrace{4n_{\text{tr}}^{-1}\mathbb{E}\left[\delta^\top \widehat{D}^{-1}\Sigma\widehat{D}^{-1}\delta\right]}_{(II)''}.\end{aligned}$$

For $(I)''$, applying Lemma F.0.7 yields

$$\begin{aligned}(I)'' &= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{i'=1}^{d}\sum_{j'=1}^{d}\delta_i\delta_j\delta_{i'}\delta_{j'}\mathbb{E}\left[\frac{1}{s_i^2 s_j^2 s_{i'}^2 s_{j'}^2}\right] \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{i'=1}^{d}\sum_{j'=1}^{d}\delta_i\delta_j\delta_{i'}\delta_{j'}\left[\frac{1}{\sigma_i^2\sigma_j^2\sigma_{i'}^2\sigma_{j'}^2} + O(n^{-1})\right] \\ &= \left(\delta^\top D^{-1}\delta\right)^2 + \left(\delta^\top\delta\right)^2 \cdot O(n^{-1}).\end{aligned}$$

Similarly, for $(II)''$, Lemma F.0.7 yields

$$(II)'' \;=\; 4n_{\text{tr}}^{-1}\delta^\top D^{-1}\Sigma D^{-1}\delta + \delta^\top\Sigma\delta\cdot O(n^{-2}).$$

Since the eigenvalues of Σ are uniformly bounded and $\delta^\top\Sigma^{-1}\delta = O(n^{-1/2})$, the variance is bounded by

$$\begin{aligned}\text{Var}[V_{0,\widehat{D}^{-1}} + V_{1,\widehat{D}^{-1}}] &= \left(\delta^\top D^{-1}\delta\right)^2\cdot O(n^{-1}) + \left(\delta^\top\delta\right)^2\cdot O(n^{-1}) \\ &\quad + 4n_{\text{tr}}^{-1}\delta^\top D^{-1}\Sigma D^{-1}\delta + \delta^\top\Sigma\delta\cdot O(n^{-2}) \\ &= O\left(\frac{1}{n^{3/2}}\right).\end{aligned}$$

This verifies the third part.

• **Concluding the proof.**

Consider a bivariate function $f(v,u) = \Phi(v/\sqrt{u})$. Then by the Taylor expansion of $f(v,u)$ around $(\Psi_{D^{-1},n,d}, \Lambda_{D^{-1},n,d})$ together with the results in Part 1 and Part 2, we have

$$\begin{aligned}&\mathcal{E}_{0,\widehat{D}^{-1}} = f(V_{0,\widehat{D}^{-1}}, U_{\widehat{D}^{-1}}) \\ =\;& \Phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right) + \phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right)\frac{1}{\sqrt{\Lambda_{D^{-1},n,d}}}(V_{0,\widehat{D}^{-1}} - \Psi_{D^{-1},n,d}) \\ &-\phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right)\frac{\Psi_{D^{-1},n,d}}{(\Lambda_{D^{-1},n,d})^{3/2}}(U_{\widehat{D}^{-1}} - \Lambda_{D^{-1},n,d}) + O_P\left(n^{-1}\right),\end{aligned}$$

where $\phi(\cdot)$ is the density function of $N(0,1)$. Similarly,

$$\begin{aligned}&\mathcal{E}_{1,\widehat{D}^{-1}} = f(V_{1,\widehat{D}^{-1}}, U_{\widehat{D}^{-1}}) \\ =\;& \Phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right) + \phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right)\frac{1}{\sqrt{\Lambda_{D^{-1},n,d}}}(V_{1,\widehat{D}^{-1}} - \Psi_{D^{-1},n,d}) \\ &-\phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right)\frac{\Psi_{D^{-1},n,d}}{(\Lambda_{D^{-1},n,d})^{3/2}}(U_{\widehat{D}^{-1}} - \Lambda_{D^{-1},n,d}) + O_P\left(n^{-1}\right).\end{aligned}$$

Combining these approximations with the result in Part 3,

$$\begin{aligned}&\frac{\mathcal{E}_{0,\widehat{D}^{-1}} + \mathcal{E}_{1,\widehat{D}^{-1}}}{2} \\ =\;& \Phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right) + \phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right)\frac{1}{\sqrt{\Lambda_{D^{-1},n,d}}}\left(\frac{V_{0,\widehat{D}^{-1}} + V_{1,\widehat{D}^{-1}}}{2} - \Psi_{D^{-1},n,d}\right)\end{aligned}$$

$$
\begin{aligned}
&-\phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right)\frac{\Psi_{D^{-1},n,d}}{(\Lambda_{D^{-1},n,d})^{3/2}}(U_{\widehat{D}^{-1}}-\Lambda_{D^{-1},n,d})+O_P\left(n^{-1}\right)\\
&=\ \Phi\left(\frac{\Psi_{D^{-1},n,d}}{\sqrt{\Lambda_{D^{-1},n,d}}}\right)+O_P\left(\frac{1}{n^{3/4}}\right).
\end{aligned}
$$

This completes the proof of Lemma F.0.5.

F.3.9 Proof of Theorem 7.5

The proof of Theorem 7.5 basically follows the same lines of arguments as in the proof of Theorem 7.3 under the given assumptions. However we note that the proof of Theorem 7.3 relies on Lemma F.0.1, which is tailored to the normality assumption. Hence, in order to complete the proof, we need to verify that the parts that build on Lemma F.0.1 are also valid for elliptical distributions. More specifically there are two parts that depend on Lemma F.0.1: (i) the approximations of $V_{0,A}$, $V_{1,A}$ and $V_{0,A}+V_{1,A}$ given in (F.16) and (ii) the approximation of U_A given in (F.21). In the rest of the proof, we prove that these approximations are still valid for elliptical distributions.

• **Moments of elliptical distributions.** Let us start with some useful moment expressions of an elliptical random vector.

Lemma F.0.8 (Chapter 3.2 of Mathai et al. (2012)). *Suppose that $Z=(Z_1,\ldots,Z_d)^\top\in\mathbb{R}^d$ has a multivariate elliptical distribution with parameters (μ,S,ξ) where $\mu=(\mu_1,\ldots,\mu_d)^\top$ and $[\Sigma]_{jk}=-2\xi'(0)[S]_{jk}=\sigma_{jk}$ for $j,k=1,\ldots,d$ such that*

$$
\mathbb{E}\big[e^{it^\top Z}\big]=e^{it^\top\mu}\xi\big(t^\top St\big)\quad \textit{for all } t\in\mathbb{R}^d.
$$

Then we have

1. $\mathbb{E}[Z_j]=\mu_j$,
2. $\mathbb{E}[Z_jZ_k]=\mu_j\mu_k+\sigma_{jk}$,
3. $\mathbb{E}[Z_jZ_kZ_l]=\mu_j\mu_k\mu_l+\mu_j\sigma_{lk}+\mu_k\sigma_{jl}+\mu_l\sigma_{jk}$.

Moreover for a symmetric matrix A, we have

1. $\mathbb{E}[Z^\top AZ]=\mu^\top A\mu+\mathrm{tr}(A\Sigma)$,
2. $Var[Z^\top AZ]=4\mu^\top A\Sigma A\mu+\zeta_{\mathrm{kurt}}\{\mathrm{tr}(A\Sigma)\}^2+2(\zeta_{\mathrm{kurt}}+1)\mathrm{tr}\{(A\Sigma)^2\}$,

where

$$\zeta_{\mathrm{kurt}} = \frac{\xi''(0)}{\{\xi'(0)\}^2} - 1 = \frac{\mathbb{E}\big[\{(Z-\mu)^\top \Sigma^{-1}(Z-\mu)\}^2\big]}{d(d+2)} - 1.$$

Note that when Z has an multivariate normal distribution, the kurtosis parameter becomes $\zeta_{\mathrm{kurt}} = 0$ and the above result coincides with Lemma F.0.1.

• **Part 1. Approximation (F.16)** Leveraging Lemma F.0.8, we first prove that the approximations of $V_{0,A} = \widehat{\delta}^\top A(\mu_0 - \widehat{\mu}_{\mathrm{pool}}), V_{1,A} = \widehat{\delta}^\top A(\widehat{\mu}_{\mathrm{pool}} - \mu_1)$ and $V_{0,A} + V_{1,A} = \widehat{\delta}^\top A(\mu_0 - \mu_1)$ in (F.16) hold true for elliptical distributions under **(A7)**. By assuming $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$, it is straightforward to see that the expected values of these quantities are

$$\mathbb{E}[V_{0,A}] = \mathbb{E}[V_{1,A}] = -\frac{1}{2}\delta^\top A\delta \quad \text{and}$$

$$\mathbb{E}[V_{0,A} + V_{1,A}] = -\delta^\top A\delta.$$

Turning to the variances, we shall prove that $\mathrm{Var}[V_{0,A}] = O(n^{-1})$, $\mathrm{Var}[V_{1,A}] = O(n^{-1})$ and $\mathrm{Var}[V_{0,A}+V_{1,A}] = O(n^{-3/2})$, which in turn yields the claim (F.16). Focusing on the variance of $V_{0,A}$ and using $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$, we see that

$$\begin{aligned}
\mathrm{Var}[V_{0,A}] \;=\; & \frac{1}{n_{0,\mathrm{tr}}^4}\mathrm{Var}\Bigg[\sum_{i=1}^{n_{0,\mathrm{tr}}}\Big\{(X_i - Y_i)^\top A\Big(\mu_0 - \frac{1}{2}X_i - \frac{1}{2}Y_i\Big)\Big\} \\
& \qquad + \sum_{1\le i\neq j\le n_{0,\mathrm{tr}}}\Big\{(X_i - Y_i)^\top A\Big(\mu_0 - \frac{1}{2}X_j - \frac{1}{2}Y_j\Big)\Big\}\Bigg] \\
\le\; & \underbrace{\frac{2}{n_{0,\mathrm{tr}}^4}\mathrm{Var}\Bigg[\sum_{i=1}^{n_{0,\mathrm{tr}}}\Big\{(X_i - Y_i)^\top A\Big(\mu_0 - \frac{1}{2}X_i - \frac{1}{2}Y_i\Big)\Big\}\Bigg]}_{(I)} \\
& + \underbrace{\frac{2}{n_{0,\mathrm{tr}}^4}\mathrm{Var}\Bigg[\sum_{1\le i\neq j\le n_{0,\mathrm{tr}}}\Big\{(X_i - Y_i)^\top A\Big(\mu_0 - \frac{1}{2}X_j - \frac{1}{2}Y_j\Big)\Big\}\Bigg]}_{(II)}
\end{aligned}$$

where the last inequality follows by $\mathrm{Var}[X+Y] \le 2\mathrm{Var}[X] + 2\mathrm{Var}[Y]$. For the first term (I), since we assume $\mathcal{X}_0^{n_0}$ and $\mathcal{Y}_0^{n_1}$ are mutually independent, we have

$$(I) \;=\; \frac{1}{2n_{0,\mathrm{tr}}^3}\mathrm{Var}\big\{(X_1 - Y_1)^\top A(2\mu_0 - X_1 - Y_1)\big\}$$

$$= \frac{1}{n_{0,\mathrm{tr}}^3}\big[2(\zeta_{\mathrm{kurt}}+1)\mathrm{tr}\big\{(A\Sigma)^2\big\} + \zeta_{\mathrm{kurt}}\big\{\mathrm{tr}(A\Sigma)\big\}^2 + 2\delta^\top A\Sigma A\delta\big],$$

where the second equality follows by straightforward calculation using Lemma F.0.8. Thus under the given conditions, we have established that $(I) = O(n^{-1})$. For the second term (II), by expanding the variance of the sum of random variables, we see that

$$\begin{aligned}
(II) \;&=\; O(n^{-2})\cdot \mathrm{Cov}\big\{(X_1-Y_1)^\top A(2\mu_0-X_2-Y_2),\ (X_1-Y_1)^\top A(2\mu_0-X_2-Y_2)\big\} \\
&\quad + O(n^{-2})\cdot \mathrm{Cov}\big\{(X_1-Y_1)^\top A(2\mu_0-X_2-Y_2),\ (X_2-Y_2)^\top A(2\mu_0-X_1-Y_1)\big\} \\
&\quad + O(n^{-1})\cdot \mathrm{Cov}\big\{(X_1-Y_1)^\top A(2\mu_0-X_2-Y_2),\ (X_2-Y_2)^\top A(2\mu_0-X_3-Y_3)\big\} \\
&\quad + O(n^{-1})\cdot \mathrm{Cov}\big\{(X_1-Y_1)^\top A(2\mu_0-X_2-Y_2),\ (X_3-Y_3)^\top A(2\mu_0-X_1-Y_1)\big\} \\
&\overset{\mathrm{def}}{=}\; O(n^{-2})\cdot(II_1) + O(n^{-2})\cdot(II_2) + O(n^{-1})\cdot(II_3) + O(n^{-1})\cdot(II_4).
\end{aligned}$$

Again, leveraging Lemma F.0.8, it can be seen that

$$\begin{aligned}
(II_1) \;&=\; 4\delta^\top A\Sigma A\delta + 4\mathrm{tr}\{(A\Sigma)^2\}, \\
(II_2) \;&=\; -4\delta^\top A\Sigma A\delta + 4\mathrm{tr}\{(A\Sigma)^2\}, \\
(II_3) \;&=\; 2\delta^\top A\Sigma A\delta \quad \text{and} \\
(II_4) \;&=\; 2\delta^\top A\Sigma A\delta.
\end{aligned}$$

Thus, under the given conditions, we can conclude that $\mathrm{Var}[V_{0,A}] = O(n^{-1})$. By symmetry we similarly have $\mathrm{Var}[V_{1,A}] = O(n^{-1})$. For the last quantity $V_{0,A} + V_{1,A}$,

$$\begin{aligned}
\mathrm{Var}[V_{0,A}+V_{1,A}] \;&=\; \mathrm{Var}[\widehat{\delta}^\top A(\mu_0-\mu_1)] = \frac{1}{n_{0,\mathrm{tr}}^2}\mathrm{Var}[(X_1-Y_1)^\top A(\mu_0-\mu_1)] \\
&=\; \frac{2}{n_{0,\mathrm{tr}}^2}\delta^\top A\Sigma A\delta = O(n^{-3/2}).
\end{aligned}$$

Combining the pieces together proves the validity of the approximations (F.16).

• **Part 2. Approximation (F.21)** Recall that $U_A = \widehat{\delta}^\top A\Sigma A\widehat{\delta}$ and it is relatively straightforward to compute the expectation under $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$ as

$$\mathbb{E}[U_A] \;=\; \delta^\top A\Sigma A\delta + \frac{2}{n_{0,\mathrm{tr}}^2}\mathrm{tr}\big\{(A\Sigma)^2\big\}.$$

Therefore it is enough to show that the variance of U_A is $O(n^{-1})$, which in turns proves the claim (F.21). Similarly as before in part 1, we can upper bound the variance of U_A by

$$\begin{aligned}\mathrm{Var}[U_A] \;\le\; & \underbrace{\frac{2}{n_{0,\mathrm{tr}}^3}\mathrm{Var}\big\{(X_1-Y_1)^\top A\Sigma A(X_1-Y_1)\big\}}_{(I)} \\ & + \underbrace{\frac{2}{n_{0,\mathrm{tr}}^4}\mathrm{Var}\Bigg\{\sum_{1\le i\ne j\le n_{0,\mathrm{tr}}}(X_i-Y_i)^\top A\Sigma A(X_j-Y_j)\Bigg\}}_{(II)}.\end{aligned}$$

For the first term (I), we observe that by the independence between X_1 and Y_1, the characteristic function of $Z_1 \stackrel{\mathrm{def}}{=} X_1 - Y_1$ is

$$\mathbb{E}\big[e^{it^\top Z_1}\big] = e^{it^\top\delta}\xi^2(t^\top St).$$

In other words, Z_1 has an elliptical distribution with parameters (δ, S, ξ^2). Also the corresponding covariance matrix and the kurtosis parameter of Z_1 are 2Σ and $\zeta_{\mathrm{kurt}}/2$, respectively. Then using Lemma F.0.8 yields

$$\begin{aligned}(I) \;&=\; \frac{4}{n_{0,\mathrm{tr}}^3}\big[4\delta^\top A\Sigma A\Sigma A\Sigma A\delta + \zeta_{\mathrm{kurt}}\{\mathrm{tr}(A\Sigma A\Sigma)\}^2 + 2(\zeta_{\mathrm{kurt}}+2)\mathrm{tr}\{(A\Sigma)^4\}\big] \\ &=\; O(n^{-1}).\end{aligned}$$

Let Z_2, Z_3 be independent copies of Z_1. Then for the second term (II),

$$\begin{aligned}(II) \;&=\; O(n^{-2})\cdot\underbrace{\mathrm{Cov}\big\{Z_1^\top A\Sigma AZ_2,\; Z_1^\top A\Sigma AZ_2\big\}}_{(II_1)} \\ &\quad + O(n^{-1})\cdot\underbrace{\mathrm{Cov}\big\{Z_1^\top A\Sigma AZ_2,\; Z_1^\top A\Sigma AZ_3\big\}}_{(II_2)}.\end{aligned}$$

Building on Lemma F.0.8, it can be shown that

$$\begin{aligned}(II_1) \;&=\; 4\delta^\top A\Sigma A\Sigma A\Sigma A\delta + 4\mathrm{tr}\{(A\Sigma)^4\} \\ (II_2) \;&=\; 2\delta^\top A\Sigma A\Sigma A\Sigma A\delta.\end{aligned}$$

Therefore the second term also satisfies $(II) = O(n^{-1})$, which verifies the claim (F.21). This completes the proof of Theorem 7.5.

F.3.10 Proof of Proposition 7.3

We let denote the conditional expectations of $\widehat{E}_0^S(\widehat{C})$ and $\widehat{E}_1^S(\widehat{C})$ given the training set by

$$\mathcal{E}_0(\widehat{C}) \stackrel{\text{def}}{=} \Pr_{Z\sim\mathbb{P}_0}\left(\widehat{C}(Z) = 1 \mid \mathcal{X}_1^{n_{0,\text{tr}}}, \mathcal{Y}_1^{n_{1,\text{tr}}}\right) \quad \text{and}$$

$$\mathcal{E}_1(\widehat{C}) \stackrel{\text{def}}{=} \Pr_{Z\sim\mathbb{P}_1}\left(\widehat{C}(Z) = 0 \mid \mathcal{X}_1^{n_{0,\text{tr}}}, \mathcal{Y}_1^{n_{1,\text{tr}}}\right).$$

For the rest of the proof, we omit the dependence of $\widehat{C}$ on the classification errors to simplify the notation.

Now, since $\widehat{E}_0^S$ and $\widehat{E}_1^S$ are uniformly bounded, the convergence in probability implies that the convergence in moment. Hence we have that $\mathbb{E}\big[\widehat{E}_0^S\big] \to E_0$ and $\mathbb{E}\big[\widehat{E}_1^S\big] \to E_1$, which implies $\mathcal{E}_0 \xrightarrow{p} E_0$ and $\mathcal{E}_1 \xrightarrow{p} E_1$ using Markov's inequality. Consequently,

$$\frac{\widehat{E}_0^S(1-\widehat{E}_0^S)/n_{0,\text{te}} + \widehat{E}_1^S(1-\widehat{E}_1^S)/n_{1,\text{te}}}{\mathcal{E}_0(1-\mathcal{E}_0)/n_{0,\text{te}} + \mathcal{E}_1(1-\mathcal{E}_1)/n_{1,\text{te}}} \xrightarrow{p} 1. \tag{F.39}$$

Suppose that the null hypothesis is true. Then under the given conditions, following the same lines of the proof of Proposition 7.2 yields

$$\frac{2\widehat{E}^S - 1}{\sqrt{\mathcal{E}_0(1-\mathcal{E}_0)/n_{0,\text{te}} + \mathcal{E}_1(1-\mathcal{E}_1)/n_{1,\text{te}}}} \xrightarrow{d} N(0,1),$$

where we use the fact that $\mathcal{E}_0 + \mathcal{E}_1 = 1$ under the null hypothesis. It is worth mentioning that Proposition 7.2 also requires **(A1)**, **(A2)**, **(A5)** and **(A6)**. These assumptions are made to show that $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ are asymptotically bounded below by 0 and above by 1, which are guaranteed by the assumption **(A9)** under the current setting.

Next Slutsky's theorem together with the observation (F.39) further shows that

$$\frac{2\widehat{E}^S - 1}{\sqrt{\widehat{E}_0^S(1-\widehat{E}_0^S)/n_{0,\text{te}} + \widehat{E}_1^S(1-\widehat{E}_1^S)/n_{1,\text{te}}}} \xrightarrow{d} N(0,1).$$

Therefore $\varphi_{\widehat{C},\text{Asymp}}$ asymptotically controls the type-1 error rate under the given conditions. In terms of power, the assumption **(A9)** guarantees that $2\widehat{E}^S - 1 \xrightarrow{p} -2\epsilon < 0$ and

$$\widehat{E}_0^S(1-\widehat{E}_0^S)/n_{0,\text{te}} + \widehat{E}_1^S(1-\widehat{E}_1^S)/n_{1,\text{te}} \xrightarrow{p} 0.$$

Building on this observation, we have under the alternative that

$$\mathbb{E}_{H_1}\big[\varphi_{\widehat{C},\text{Asymp}}\big]$$

$$
\begin{aligned}
&= \mathbb{P}_{H_1}\left[\frac{2\widehat{E}^S - 1}{\sqrt{\widehat{E}_0^S(1-\widehat{E}_0^S)/n_{0,\mathrm{te}} + \widehat{E}_1^S(1-\widehat{E}_1^S)/n_{1,\mathrm{te}}}} < -z_\alpha\right] \\
&= \mathbb{P}_{H_1}\left[2\widehat{E}^S - 1 < -z_\alpha\sqrt{\widehat{E}_0^S(1-\widehat{E}_0^S)/n_{0,\mathrm{te}} + \widehat{E}_1^S(1-\widehat{E}_1^S)/n_{1,\mathrm{te}}}\right] \\
&\to 1,
\end{aligned}
$$

which proves consistency of the asymptotic test.

F.3.11 Proof of Theorem 7.6

As mentioned in the main text, both half- and entire-permutation methods yield a valid level α test (see, e.g., Theorem 1 of Hemerik and Goeman, 2018b). Hence we focus on proving consistency of the resulting test under the given conditions. To ease notation, we drop the dependence of $\widehat{C}$ on the sample-splitting errors throughout this proof.

Let us consider all possible permutations first, that is $m! \overset{\text{def}}{=} n_{\mathrm{te}}!$ for method 1 and $m! \overset{\text{def}}{=} n!$ for method 2, and denote the sample-splitting errors (or $1-$ accuracies) by $\widehat{E}^{S,1}, \ldots, \widehat{E}^{S,m!}$ computed based on each permutation. We then let $\widetilde{E}^{S,1}, \ldots, \widetilde{E}^{S,P}$ be P independent samples from $\widehat{E}^{S,1}, \ldots, \widehat{E}^{S,m!}$ without replacement. Then the permutation test can be equivalently written as

$$
\varphi_{\widehat{C},\mathrm{Perm}} = \mathbb{I}\left[\frac{1}{P}\sum_{i=1}^{P}\mathbb{I}\left(\widehat{E}^S < \widetilde{E}^{S,i}\right) \geq 1 - \alpha_P\right], \tag{F.40}
$$

where $1 - \alpha_P \overset{\text{def}}{=} \lceil (1-\alpha)(1+P)\rceil / P \to 1-\alpha$ as $P \to \infty$. We note that in order for the test (F.40) to have power, $1-\alpha_P$ should be less than one (otherwise the test function is always zero), which requires the condition $P > (1-\alpha)/\alpha$.

Let us denote the α_P quantile of $\widetilde{E}^{S,1}, \ldots, \widetilde{E}^{S,P}$ by q_{α_P}. Using the representation (F.40), it can be verified that if the test statistic is less than this quantile, i.e. $\widehat{E}^S < q_{\alpha_P}$, then the permutation test is equal to one, i.e. $\varphi_{\widehat{C},\mathrm{Perm}} = 1$. This fact implies that if $\mathbb{I}(\widehat{E}^S < q_{\alpha_P})$ is a consistent test, then the permutation test is also consistent. Therefore it is enough to work with $\mathbb{I}(\widehat{E}^S < q_{\alpha_P})$ and show that it is consistent.

A high-level proof strategy is as follows. By the assumption, $\widehat{E}^S$ converges in probability to a constant strictly less than $1/2 - \epsilon/2$ under the alternative. Therefore the proof is complete if we show that a lower bound for q_{α_P} converges to a constant that is strictly larger than $1/2 - \epsilon/2$. To do so, we let $\mathfrak{n}$ be a random variable uniformly distributed over $\{1, \ldots, P\}$ and write the distribution of $\mathfrak{n}$ by $\mathbb{P}_{\mathfrak{n}}$ (conditional on everything else) and the expectation with respect to $\mathbb{P}_{\mathfrak{n}}$ by $\mathbb{E}_{\mathfrak{n}}$.

For a given $t \in (0, 1/2)$, applying Markov's inequality yields

$$\begin{aligned}\mathbb{P}_{\mathfrak{n}}\Big(\widetilde{E}^{S,\mathfrak{n}} < t\Big) &= \mathbb{P}_{\mathfrak{n}}\Big(- \widetilde{E}^{S,\mathfrak{n}} + 1/2 > -t + 1/2\Big) \\ &\leq \mathbb{P}_{\mathfrak{n}}\Big(\big|\widetilde{E}^{S,\mathfrak{n}} - 1/2\big| > -t + 1/2\Big) \\ &\leq \frac{\mathbb{E}_{\mathfrak{n}}\big[\big(\widetilde{E}^{S,\mathfrak{n}} - 1/2\big)^2\big]}{(1/2 - t)^2}.\end{aligned}$$

Now by setting the right-hand side to be α_P, we know that the quantile q_{α_P} is lower bounded by

$$q_{\alpha_P} \geq \frac{1}{2} - \sqrt{\frac{1}{\alpha_P}\mathbb{E}_{\mathfrak{n}}\big[\big(\widetilde{E}^{S,\mathfrak{n}} - 1/2\big)^2\big]}.$$

Here the expected value of the squared difference is

$$\mathbb{E}_{\mathfrak{n}}\big[\big(\widetilde{E}^{S,\mathfrak{n}} - 1/2\big)^2\big] = \frac{1}{P}\sum_{i=1}^{P}\Big(\widetilde{E}^{S,i} - 1/2\Big)^2. \tag{F.41}$$

In the rest of the proof, we show that the above quantity converges in probability to zero as $n \to \infty$ for both method 1 and method 2. Hence the quantile q_{α_P} is lower bounded by $1/2 - \epsilon/2$ in the limit as claimed.

• Method 1 (Half-permutation test).

To start with method 1, we let $\mathfrak{m}$ be a random variable uniformly distributed over $\{1, \ldots, n_{\mathrm{te}}!\}$ and write the expectation and the variance over $\mathfrak{m}$ (conditional on everything else) by $\mathbb{E}_{\mathfrak{m}}$ and $\mathbb{V}_{\mathfrak{m}}$, respectively. We note that for each $i \in \{1, \ldots, P\}$, $\widetilde{E}^{S,i}$ has the same distribution as $\widehat{E}^{S,\mathfrak{m}}$ and that the expected value of $\widehat{E}^{S,\mathfrak{m}}$ is calculated as

$$\mathbb{E}_{\mathfrak{m}}\big[\widehat{E}^{S,\mathfrak{m}}\big] = \frac{1}{n_{\mathrm{te}}!}\sum_{i=1}^{n_{\mathrm{te}}!}\widehat{E}^{S,i} = \frac{1}{2}. \tag{F.42}$$

Therefore the squared difference (F.41) is an unbiased estimator of the variance of $\widehat{E}^{S,\mathfrak{m}}$.

We next upper bound the variance of $\widehat{E}^{S,\mathfrak{m}}$. To do so, let us write the test set by

$$\{X_{1+n_{0,\mathrm{tr}}}, \ldots, X_{n_0}, Y_{1+n_{1,\mathrm{tr}}}, \ldots, Y_{n_1}\} \stackrel{\mathrm{def}}{=} \{Z_1, \ldots, Z_{n_{\mathrm{te}}}\}.$$

Notice that for each $\mathfrak{m}$, there exists the corresponding permutation of $\{1,\ldots,n_{\text{te}}\}$, denoted by $\omega^{\mathfrak{m}} \stackrel{\text{def}}{=} \{\omega_1^{\mathfrak{m}},\ldots,\omega_{n_{\text{te}}}^{\mathfrak{m}}\}$, such that the test statistic $\widehat{E}^{S,\mathfrak{m}}$ can be written as

$$\widehat{E}^{S,\mathfrak{m}} \;=\; \underbrace{\frac{1}{2n_{0,\text{te}}}\sum_{i=1}^{n_{0,\text{te}}}\mathbb{I}\Big[\widehat{C}(Z_{\omega_i^{\mathfrak{m}}})=1\Big]}_{(I)} + \underbrace{\frac{1}{2n_{1,\text{te}}}\sum_{i=1}^{n_{1,\text{te}}}\mathbb{I}\Big[\widehat{C}(Z_{\omega_{i+n_{0,\text{te}}}^{\mathfrak{m}}})=0\Big]}_{(II)}.$$

The variance of the first term (I) is

$$\begin{aligned}\text{Var}_{\mathfrak{m}}[(I)] \;&=\; \frac{1}{4n_{0,\text{te}}^2}\sum_{i=1}^{n_{0,\text{te}}}\text{Var}_{\mathfrak{m}}\Big\{\mathbb{I}\Big[\widehat{C}(Z_{\omega_i^{\mathfrak{m}}})=1\Big]\Big\}\\ &\quad+\frac{1}{4n_{0,\text{te}}^2}\sum_{1\le i\neq j\le n_{0,\text{te}}}\text{Cov}_{\mathfrak{m}}\Big\{\mathbb{I}\Big[\widehat{C}(Z_{\omega_i^{\mathfrak{m}}})=1\Big],\;\mathbb{I}\Big[\widehat{C}(Z_{\omega_j^{\mathfrak{m}}})=1\Big]\Big\},\end{aligned}$$

where the individual variance and covariance terms are given as

$$\text{Var}_{\mathfrak{m}}\Big\{\mathbb{I}\Big[\widehat{C}(Z_{\omega_i^{\mathfrak{m}}})=1\Big]\Big\} = \;\frac{1}{n_{\text{te}}}\sum_{i=1}^{n_{\text{te}}}\mathbb{I}\Big[\widehat{C}(Z_i)=1\Big]\cdot\Big\{1-\frac{1}{n_{\text{te}}}\sum_{i=1}^{n_{\text{te}}}\mathbb{I}\Big[\widehat{C}(Z_i)=1\Big]\Big\}\le\;1$$

and

$$\begin{aligned}&\text{Cov}_{\mathfrak{m}}\Big\{\mathbb{I}\Big[\widehat{C}(Z_{\omega_i^{\mathfrak{m}}})=1\Big],\;\mathbb{I}\Big[\widehat{C}(Z_{\omega_j^{\mathfrak{m}}})=1\Big]\Big\}\\ =\;&\frac{1}{n_{\text{te}}(n_{\text{te}}-1)}\sum_{1\le i\neq j\le n_{\text{te}}}\mathbb{I}\Big[\widehat{C}(Z_i)=1\Big]\cdot\mathbb{I}\Big[\widehat{C}(Z_j)=1\Big]-\Big\{\frac{1}{n_{\text{te}}}\sum_{i=1}^{n_{\text{te}}}\mathbb{I}\Big[\widehat{C}(Z_i)=1\Big]\Big\}^2\\ \le\;&0.\end{aligned}$$

Hence the variance of (I) is bounded by $\text{Var}_{\mathfrak{m}}[(I)]\le 1/(4n_{0,\text{te}})$ and similarly one can show that $\text{Var}_{\mathfrak{m}}[(II)]\le 1/(4n_{1,\text{te}})$. Now applying the basic inequality $\text{Var}(X+Y)\le 2\text{Var}(X)+2\text{Var}(Y)$ yields

$$\text{Var}_{\mathfrak{m}}\big[\widehat{E}^{S,\mathfrak{m}}\big]\le\frac{1}{2n_{0,\text{te}}}+\frac{1}{2n_{1,\text{te}}}. \tag{F.43}$$

This in turn implies that $(\widetilde{E}^{S,i}-1/2)^2\xrightarrow{p}0$ as $n\to\infty$ for any $i\in\{1,\ldots,P\}$ and thus

$$\frac{1}{P}\sum_{i=1}^{P}\Big(\widetilde{E}^{S,i}-1/2\Big)^2\xrightarrow{p}0\quad\text{as } n\to\infty. \tag{F.44}$$

This completes the proof for method 1.

• **Method 2 (Entire-permutation test).**

Next we show that the squared difference (F.41) converges to zero in probability for method 2. We first note that the half-permutation procedure can be understood as the entire-permutation procedure conditional on the first n_{tr} permutation labels. From this perspective, $\mathbb{E}_{\mathfrak{m}}$ and $\mathrm{Var}_{\mathfrak{m}}$ are the conditional expectation and the conditional variance of the permuted test statistic given the first n_{tr} permutation labels. More specifically we let $\mathfrak{m}^*$ be a random variable uniformly distributed over $\{1, \ldots, n!\}$ and write the distribution of $\mathfrak{m}^*$ by $\mathbb{P}_{\mathfrak{m}^*}$ (conditional on everything else) and the expectation and the variance with respect to $\mathbb{P}_{\mathfrak{m}^*}$ by $\mathbb{E}_{\mathfrak{m}^*}$ and $\mathbb{V}_{\mathfrak{m}^*}$, respectively. Then for each $\mathfrak{m}^*$, there exists the corresponding permutation of $\{1, \ldots, n\}$, denoted by $\omega^{\mathfrak{m}^*} \stackrel{\mathrm{def}}{=} \{\omega_1^{\mathfrak{m}^*}, \ldots, \omega_n^{\mathfrak{m}^*}\}$, such that the permuted test statistic can be expressed as a function of $\omega^{\mathfrak{m}^*}$ as

$$\widehat{E}^{S,\mathfrak{m}^*} \stackrel{\mathrm{def}}{=} \widehat{E}^{S,\mathfrak{m}^*}(Z_{\omega_1^{\mathfrak{m}^*}}, \ldots, Z_{\omega_n^{\mathfrak{m}^*}}),$$

where $\{Z_1, \ldots, Z_n\}$ are the pooled samples denoted by

$$\{Z_1, \ldots, Z_n\} \stackrel{\mathrm{def}}{=} \{X_1, \ldots, X_{n_{0,\mathrm{tr}}}, Y_1, \ldots, Y_{n_{1,\mathrm{tr}}}, X_{1+n_{0,\mathrm{tr}}}, \ldots, X_{n_0}, Y_{1+n_{1,\mathrm{tr}}}, \ldots, Y_{n_1}\}.$$

Following the same reasoning in (F.42), it can be seen that the conditional expectation of $\widehat{E}^{S,\mathfrak{m}^*}$ given the first n_{tr} components of $\omega^{\mathfrak{m}^*}$ is always equal to half, that is

$$\mathbb{E}_{\mathfrak{m}^*}\left[\widehat{E}^{S,\mathfrak{m}^*} \,\middle|\, \omega_1^{\mathfrak{m}^*}, \ldots, \omega_{n_{\mathrm{tr}}}^{\mathfrak{m}^*}\right] = \frac{1}{2}.$$

Hence applying the law of total expectation yields that the unconditional expectation is also equal to half. Next we use the law of total variance and observe that

$$\begin{aligned}
\mathbb{V}_{\mathfrak{m}^*}\left[\widehat{E}^{S,\mathfrak{m}^*}\right] &= \mathbb{V}_{\mathfrak{m}^*}\left[\mathbb{E}_{\mathfrak{m}^*}\left\{\widehat{E}^{S,\mathfrak{m}^*} \,\middle|\, \omega_1^{\mathfrak{m}^*}, \ldots, \omega_{n_{\mathrm{tr}}}^{\mathfrak{m}^*}\right\}\right] \\
&\quad + \mathbb{E}_{\mathfrak{m}^*}\left[\mathrm{Var}_{\mathfrak{m}^*}\left\{\widehat{E}^{S,\mathfrak{m}^*} \,\middle|\, \omega_1^{\mathfrak{m}^*}, \ldots, \omega_{n_{\mathrm{tr}}}^{\mathfrak{m}^*}\right\}\right] \\
&= \mathbb{E}_{\mathfrak{m}^*}\left[\mathrm{Var}_{\mathfrak{m}^*}\left\{\widehat{E}^{S,\mathfrak{m}^*} \,\middle|\, \omega_1^{\mathfrak{m}^*}, \ldots, \omega_{n_{\mathrm{tr}}}^{\mathfrak{m}^*}\right\}\right] \;\leq\; \frac{1}{2n_{0,\mathrm{te}}} + \frac{1}{2n_{1,\mathrm{te}}},
\end{aligned}$$

where the last inequality can be similarly proved as in the bound (F.43). Having these two observations at hand, we know that conclusion (F.44) is also true for method 2 and thus complete the proof of Theorem 7.6.

F.4 Simulation results on sample-splitting ratio

In this section we examine the power of classification tests under the Gaussian setting by varying the splitting ratio κ for the balanced sample case. As in Section 7.10 of the main text, we set $n_0 = n_1 = d = 200$ and consider the accuracy tests $\varphi_{\Sigma^{-1}}$ and $\varphi_{\widehat{D}^{-1}}$ based on the Fisher's LDA classifier and the naive Bayes classifier,

respectively. Note that the critical values of $\varphi_{\Sigma^{-1}}$ and $\varphi_{\widehat{D}^{-1}}$ are chosen based on a normal approximation. Given $\kappa \in \{0.1, 0.2, \ldots, 0.9\}$, the number of samples in the training set is decided by $n_{0,\mathrm{tr}} = \lfloor \kappa n_0 \rfloor$ and $n_{1,\mathrm{tr}} = \lfloor \kappa n_1 \rfloor$, which leads to $n_{0,\mathrm{te}} = n_0 - n_{0,\mathrm{tr}}$ and $n_{1,\mathrm{te}} = n_1 - n_{1,\mathrm{tr}}$.

Table F.1: Comparisons of the empirical power of classification tests by varying the sample-splitting ratio κ. The results show that the power is approximately maximized when the splitting ratio is $\kappa = 1/2$. See Appendix F.4 for details.

	Ratio κ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\delta = 0.15$	LDA	0.155	0.189	0.212	0.220	**0.224**	0.207	0.176	0.157	0.103
	Bayes	0.150	0.185	0.218	0.221	**0.222**	0.212	0.176	0.156	0.100
$\delta = 0.25$	LDA	0.437	0.616	0.691	0.714	**0.715**	0.686	0.598	0.499	0.301
	Bayes	0.406	0.613	0.682	0.710	**0.714**	0.677	0.596	0.496	0.306

The results are presented in Table F.1. It is apparent from Table F.1 that the power is maximized when the training set and the testing set are well-balanced, i.e. $\kappa = 1/2$. This coincides with our theoretical result discussed in Section 7.6. However, unlike our asymptotic power expression in (7.17) with $\lambda = 1/2$, the empirical power seems asymmetric in κ. This unexpected result might be attributed to the fact that when κ is far from $1/2$, either n_{tr} or n_{te} becomes too small to justify a normal approximation. Nevertheless, the powers in these extreme cases are less than the power in the balanced case.

Appendix G

Appendix for Chapter 8

G.1 Overview of Appendix

In this supplementary material, we provide some additional results and the technical proofs omitted in the main text. The remainder of this material is organized as follows.

- In Appendix G.2, we develop exponential inequalities for permuted linear statistics, building on the concept of negative association.
- In Appendix G.3, we provide the result that improves Theorem 8.1 based on the exponential bound in Theorem 8.3 with an extra assumption that $n_1 \asymp n_2$.
- The proof of Lemma 8.0.1 on the two moments method is provided in Appendix G.4.
- The proofs of the results on two-sample testing in Section 8.4 are presented in Appendix G.5, G.6, G.7 and G.8.
- The proofs of the results on independence testing in Section 8.5 are presented in Appendix G.9, G.10, G.11, G.12 and G.13.
- The proofs of the results on combinatorial concentration inequalities in Section 8.6 are presented in Appendix G.15 and G.16.
- The proofs of the results on adaptive tests in Section 8.7 are presented in Appendix G.17 and G.18.
- The proofs of the results on multinomial tests and Gaussian kernel tests in Section 8.8 are presented in Appendix G.19, G.20, G.21 and G.22.

G.2 Exponential inequalities for permuted linear statistics

Suppose that $\mathcal{X}_n = \{(Y_1, Z_1), \ldots, (Y_n, Z_n)\}$ is a set of bivariate random variables where $Y_i \in \mathbb{R}$ and $Z_i \in \mathbb{R}$. Following the convention, let us write the sample means of Y and Z by $\overline{Y} := n^{-1}\sum_{i=1}^n Y_i$ and $\overline{Z} := n^{-1}\sum_{i=1}^n Z_i$, respectively. The sample covariance, which measures a linear relationship between Y and Z, is given by

$$L_n := \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})(Z_i - \overline{Z}).$$

We also call L_n as a linear statistic as opposed to quadratic statistics or degenerate U-statistics considered in the main text. Let us denote the permuted linear statistic, associated with a permutation π of $\{1, \ldots, n\}$, by

$$L_n^\pi = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})(Z_{\pi_i} - \overline{Z}).$$

In this section, we provide two exponential concentration bounds for L_n^π conditional on $\mathcal{X}_n$; namely Hoeffding-type inequality (Proposition G.1) and Bernstein-type inequality (Proposition G.2). These results have potential applications in studying the power of the permutation test based on L_n and also concentration inequalities for sampling without replacement. We describe the second application in more detail in Appendix G.2.1 after we develop the results.

Related work and negative association. We should note that the same problem has been considered by several authors using Stein's method (Chatterjee, 2007), a martingale method (Chapter 4.2 of Bercu et al., 2015) and Talagrand's inequality (Albert, 2019). In fact they consider a more general linear statistic which has the form of $\sum_{i=1}^n d_{i,\pi_i}$ where $\{d_{i,j}\}_{i,j=1}^n$ is an arbitrary bivariate sequence. Thus their statistic includes L_n as a special case by letting $d_{i,\pi_i} = (Y_i - \overline{Y})(Z_{\pi_i} - \overline{Z})$. However their proofs are quite involved at the expense of being more general. Here we provide a much simpler proof with sharper constant factors by taking advantage of the decomposability of $d_{i,j}$. To this end, we utilize the concept of negative association (e.g. Joag-Dev and Proschan, 1983; Dubhashi and Ranjan, 1998), defined as follows.

Definition G.1 (Negative association)**.** *Random variables $X_1, \ldots, X_n$ are negatively associated (NA) if for every two disjoint index sets $\mathcal{I}, \mathcal{J} \subseteq \{1, \ldots, n\}$,*

$$\mathbb{E}[f(X_i, i \in \mathcal{I})g(X_j, j \in \mathcal{J})] \leq \mathbb{E}[f(X_i, i \in \mathcal{I})]\mathbb{E}[g(X_j, j \in \mathcal{J})]$$

for all functions $f : \mathbb{R}^{|\mathcal{I}|} \mapsto \mathbb{R}$ and $g : \mathbb{R}^{|\mathcal{J}|} \mapsto \mathbb{R}$ that are both non-decreasing or both non-increasing.

Let us state several useful facts about negatively associated random variables that we shall leverage to prove the main results of this section. The proofs of the given facts can be found in Joag-Dev and Proschan (1983) and Dubhashi and Ranjan (1998).

- **Fact 1.** Let $\{x_1, \ldots, x_n\}$ be a set of n real values. Suppose that $\{X_1, \ldots, X_n\}$ are random variables with the probability such that

$$\mathbb{P}(X_1 = x_{\pi_1}, \ldots, X_n = x_{\pi_n}) = \frac{1}{n!} \quad \text{for any permutation } \pi \text{ of } \{1, \ldots, n\}.$$

  Then $\{X_1, \ldots, X_n\}$ are negatively associated.

- **Fact 2.** Let $\{X_1, \ldots, X_n\}$ be negatively associated. Let $\mathcal{I}_1, \ldots, \mathcal{I}_k \subseteq \{1, \ldots, n\}$ be disjoint index sets, for some positive integer k. For $j \in \{1, \ldots, n\}$, let $h_j : \mathbb{R}^{|\mathcal{I}_k|} \mapsto \mathbb{R}$ be functions that are all non-decreasing or all non-increasing and define $Y_j = h_j(X_i, i \in \mathcal{I}_j)$. Then $\{Y_1, \ldots, Y_k\}$ are also negatively associated.

- **Fact 3.** Let $\{X_1, \ldots, X_n\}$ be negatively associated. Then for any non-decreasing functions f_i, $i \in \{1, \ldots, n\}$, we have that

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] \leq \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)]. \tag{G.1}$$

Description of the main idea. Notice that L_n^π is a function of non-i.i.d. random variables for which standard techniques relying on i.i.d. assumptions do not work directly. We avoid this difficulty by connecting L_n^π with negatively associated random variables and then applying Chernoff bound combined with the inequality (G.1). The details are as follows. For notational simplicity, let us denote

$$\{a_1, \ldots, a_n\} = \{Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y}\} \quad \text{and}$$

$$\{b_{\pi_1}, \ldots, b_{\pi_n}\} = \{Z_{\pi_1} - \overline{Z}, \ldots, Z_{\pi_n} - \overline{Z}\}.$$

To proceed, we make several important observations.

- **Observation 1.** First, since $\{b_{\pi_1}, \ldots, b_{\pi_n}\}$ has a permutation distribution, we can use Fact 1 and conclude that $\{b_{\pi_1}, \ldots, b_{\pi_n}\}$ are negatively associated.

- **Observation 2.** Second, let $\mathcal{I}_+$ be the set of indices such that $a_i > 0$ and similarly $\mathcal{I}_-$ be the set of indices such that $a_i < 0$. Since $h_i(X_i, i \in \mathcal{I}_+) = a_i X_i$ is non-decreasing function and $h_i(X_i, i \in \mathcal{I}_-) = a_i X_i$ is non-increasing functions, it can be seen that $\{a_i b_{\pi_i}\}_{i \in \mathcal{I}_+}$ and $\{a_i b_{\pi_i}\}_{i \in \mathcal{I}_-}$ are negatively

associated by Fact 2. Using this notation, the linear statistic can be written as

$$L_n^\pi = \frac{1}{n}\sum_{i\in\mathcal{I}_+} a_i b_{\pi_i} + \frac{1}{n}\sum_{i\in\mathcal{I}_-} a_i b_{\pi_i}.$$

It can be easily seen that $\mathbb{E}_\pi[b_{\pi_i}|\mathcal{X}_n] = 0$ for each i and thus $\mathbb{E}_\pi[L_n^\pi|\mathcal{X}_n] = 0$ by linearity of expectation. Hence, for $\lambda > 0$, applying the Chernoff bound yields

$$\begin{aligned}
&\mathbb{P}_\pi(L_n^\pi \geq t|\mathcal{X}_n) \\
\leq\ & e^{-\lambda t}\mathbb{E}_\pi\left[\exp\left(\lambda n^{-1}\sum_{i\in\mathcal{I}_+} a_i b_{\pi_i} + \lambda n^{-1}\sum_{i\in\mathcal{I}_-} a_i b_{\pi_i}\right)\middle|\mathcal{X}_n\right] \\
\leq\ & \frac{e^{-\lambda t}}{2}\mathbb{E}_\pi\left[\exp\left(2\lambda n^{-1}\sum_{i\in\mathcal{I}_+} a_i b_{\pi_i}\right)\middle|\mathcal{X}_n\right] + \frac{e^{-\lambda t}}{2}\mathbb{E}_\pi\left[\exp\left(2\lambda n^{-1}\sum_{i\in\mathcal{I}_-} a_i b_{\pi_i}\right)\middle|\mathcal{X}_n\right] \\
:=\ & (I) + (II),
\end{aligned}$$

where the last inequality uses the elementary inequality $xy \leq x^2/2 + y^2/2$.

- **Observation 3.** Third, based on fact that $\{a_i b_{\pi_i}\}_{i\in\mathcal{I}_+}$ and $\{a_i b_{\pi_i}\}_{i\in\mathcal{I}_-}$ are negatively associated, we may apply Fact 3 to have that

$$\begin{aligned}
(I) \ &\leq\ \frac{e^{-\lambda t}}{2}\prod_{i\in\mathcal{I}_+}\mathbb{E}_{\widetilde{b}}\big[\exp\big(2\lambda n^{-1}a_i\widetilde{b}_i\big)\big|\mathcal{X}_n\big] = \frac{e^{-\lambda t}}{2}\prod_{i=1}^n\mathbb{E}_{\widetilde{b}}\big[\exp\big(2\lambda n^{-1}a_i^+\widetilde{b}_i\big)\big|\mathcal{X}_n\big] \quad \text{and} \\
(II) \ &\leq\ \frac{e^{-\lambda t}}{2}\prod_{i\in\mathcal{I}_-}\mathbb{E}_{\widetilde{b}}\big[\exp\big(2\lambda n^{-1}a_i\widetilde{b}_i\big)\big|\mathcal{X}_n\big] = \frac{e^{-\lambda t}}{2}\prod_{i=1}^n\mathbb{E}_{\widetilde{b}}\big[\exp\big(-2\lambda n^{-1}a_i^-\widetilde{b}_i\big)\big|\mathcal{X}_n\big],
\end{aligned} \tag{G.2}$$

where $\widetilde{b}_1, \ldots, \widetilde{b}_n$ are i.i.d. random variables uniformly distributed over $\{b_1, \ldots, b_n\}$. Here a_i^+ and a_i^- represent $a_i^+ = a_i\mathbb{1}(a_i \geq 0)$ and $a_i^- = -a_i\mathbb{1}(a_i \leq 0)$ respectively.

With these upper bounds for (I) and (II) in place, we are now ready to present the main results of this section. The first one is a Hoeffding-type bound which provides a sharper constant factor than Duembgen (1998).

Proposition G.1 (Hoeffding-type bound)**.** *Let us define* $a_{\text{range}} := Y_n - Y_1$ *and* $b_{\text{range}} := Z_n - Z_1$*. Then*

$$\mathbb{P}_\pi(L_n^\pi \geq t|\mathcal{X}_n) \leq \exp\left[-\max\left\{\frac{n^2t^2}{a_{\text{range}}^2\sum_{i=1}^n b_i^2},\ \frac{n^2t^2}{b_{\text{range}}^2\sum_{i=1}^n a_i^2}\right\}\right].$$

Proof. The proof directly follows by applying Hoeffding's lemma (Hoeffding, 1963), which states that when Z has zero mean and $a \le Z \le b$,

$$\mathbb{E}[e^{\lambda Z}] \le e^{\lambda^2 (b-a)^2/8}.$$

Notice that Hoeffding's lemma yields

$$\prod_{i=1}^{n} \mathbb{E}_{\widetilde{b}}\big[\exp\big(2\lambda n^{-1} a_i^{+} \widetilde{b}_i\big)\big|\mathcal{X}_n\big] \le \exp\left\{\frac{\lambda^2 b_{\text{range}}^2}{2n^2}\sum_{i=1}^{n}(a_i^{+})^2\right\} \le \exp\left\{\frac{\lambda^2 b_{\text{range}}^2}{2n^2}\sum_{i=1}^{n} a_i^2\right\} \quad \text{and}$$

$$\prod_{i=1}^{n} \mathbb{E}_{\widetilde{b}}\big[\exp\big(2\lambda n^{-1} a_i^{-} \widetilde{b}_i\big)\big|\mathcal{X}_n\big] \le \exp\left\{\frac{\lambda^2 b_{\text{range}}^2}{2n^2}\sum_{i=1}^{n}(a_i^{-})^2\right\} \le \exp\left\{\frac{\lambda^2 b_{\text{range}}^2}{2n^2}\sum_{i=1}^{n} a_i^2\right\}.$$

Thus combining the above with the upper bounds for (I) and (II) in (G.2) yields

$$\mathbb{P}_\pi(L_n^\pi \ge t|\mathcal{X}_n) \le \exp\left\{-\lambda t + \frac{\lambda^2 b_{\text{range}}^2}{2n^2}\sum_{i=1}^{n} a_i^2\right\}.$$

By optimizing over λ on the right-hand side, we obtain that

$$\mathbb{P}_\pi(L_n^\pi \ge t|\mathcal{X}_n) \le \exp\left\{-\frac{n^2 t^2}{b_{\text{range}}^2 \sum_{i=1}^{n} a_i^2}\right\}. \tag{G.3}$$

Since $\sum_{i=1}^{n} a_{\pi_i} b_i$ and $\sum_{i=1}^{n} a_i b_{\pi_i}$ have the same permutation distribution, it also holds that

$$\mathbb{P}_\pi(L_n^\pi \ge t|\mathcal{X}_n) \le \exp\left\{-\frac{n^2 t^2}{a_{\text{range}}^2 \sum_{i=1}^{n} b_i^2}\right\}. \tag{G.4}$$

Then putting together these two bounds (G.3) and (G.4) gives the desired result. □

Note that Proposition G.1 depends on the variance of either $\{a_i\}_{i=1}^n$ or $\{b_i\}_{i=1}^n$. In the next proposition, we provide a Bernstein-type bound which depends on the variance of the bivariate sequence $\{a_i b_j\}_{i,j=1}^n$. Similar results can be found in Bercu et al. (2015) and Albert (2019) but in terms of constants, the bound below is much shaper than the previous ones.

Proposition G.2 (Bernstein-type bound)**.** *Based on the same notation in Proposition G.1, a Bernstein-type bound is provided by*

$$\mathbb{P}_\pi(L_n^\pi \ge t|\mathcal{X}_n) \le \exp\left\{-\frac{nt^2}{2n^{-2}\sum_{i,j=1}^{n} a_i^2 b_j^2 + \frac{2}{3} t \max_{1\le i,j\le n}|a_i b_j|}\right\}.$$

Proof. Once we have the upper bounds for (I) and (II) in (G.2), the remainder of the proof is routine. First it is straightforward to verify that for $|Z| \le c$, $\mathbb{E}[Z] = 0$ and $\mathbb{E}[Z^2] = \sigma^2$, we have that

$$\mathbb{E}[e^{\lambda Z}] = 1 + \sum_{k=2}^{\infty} \frac{\mathbb{E}[(\lambda Z)^k]}{k!} \le 1 + \frac{\sigma^2}{c^2} \sum_{k=2}^{\infty} \frac{\lambda^k c^k}{k!} \le \exp\left\{ \frac{\sigma^2}{c^2} \left(e^{\lambda c} - 1 - \lambda c\right) \right\}.$$

Let us write $\widehat{\sigma}_i^2 = n^{-3} a_i^2 \sum_{j=1}^n b_i^2$ and $M = n^{-1} \max_{1 \le i,j \le n} |a_i b_j|$. Then based on the above inequality, we can obtain that

$$\prod_{i=1}^n \mathbb{E}_{\widetilde{b}}\big[\exp\left(2\lambda n^{-1} a_i^+ \widetilde{b}_i\right) \big| \mathcal{X}_n \big] \le \exp\left\{ \frac{\sum_{i=1}^n \widehat{\sigma}_i^2}{M^2} \left(e^{\lambda M} - 1 - \lambda M\right) \right\} \quad \text{and}$$

$$\prod_{i=1}^n \mathbb{E}_{\widetilde{b}}\big[\exp\left(2\lambda n^{-1} a_i^- \widetilde{b}_i\right) \big| \mathcal{X}_n \big] \le \exp\left\{ \frac{\sum_{i=1}^n \widehat{\sigma}_i^2}{M^2} \left(e^{\lambda M} - 1 - \lambda M\right) \right\}.$$

Combining these two upper bounds with the result in (G.2) yields

$$\mathbb{P}_\pi \left(L_n^\pi \ge t | \mathcal{X}_n \right) \le e^{-\lambda t} \exp\left\{ \frac{\sum_{i=1}^n \widehat{\sigma}_i^2}{M^2} \left(e^{\lambda M} - 1 - \lambda M\right) \right\}.$$

By optimizing the right-hand side in terms of λ, we obtain a Bennett-type inequality

$$\mathbb{P}_\pi \left(L_n^\pi \ge t | \mathcal{X}_n \right) \le \exp\left\{ - \frac{\sum_{i=1}^n \widehat{\sigma}_i^2}{M^2} h\left(\frac{tM}{\sum_{i=1}^n \widehat{\sigma}_i^2} \right) \right\},$$

where $h(x) = (1+x)\log(1+x) - x$. Then the result follows by noting that $h(x) \ge x^2/(2 + 2x/3)$. □

In the next subsection, we apply our results to derive concentration inequalities for sampling without replacement.

G.2.1 Concentration inequalities for sampling without replacement

To establish the explicit connection to sampling without replacement, we focus on the case where Z_i is binary, say $Z_i \in \{-a, a\}$. Then the linear statistic L_n is related to the unscaled two-sample t-statistic. More specifically, let us write $n_1 = \sum_{i=1}^n \mathbb{1}(Z_i = a)$ and $n_2 = \sum_{i=1}^n \mathbb{1}(Z_i = -a)$. Additionally we use the notation $\overline{Y}_1 = n_1^{-1} \sum_{i=1}^n Y_i \mathbb{1}(Z_i = a)$ and $\overline{Y}_2 = n_2^{-1} \sum_{i=1}^n Y_i \mathbb{1}(Z_i = -a)$. Then some algebra shows that the sample covariance L_n is exactly the form of

$$L_n = 2a \frac{n_1 n_2}{n^2} (\overline{Y}_1 - \overline{Y}_2).$$

Without loss of generality, we assume $a = 1$, i.e. $Z_i \in \{-1, 1\}$. Then Proposition G.1 gives a concentration inequality for the unscaled t-statistic as

$$\mathbb{P}_\pi\left\{ \frac{2n_1n_2}{n^2}\left(\overline{Y}_{1,\pi} - \overline{Y}_{2,\pi}\right) \geq t \middle| \mathcal{X}_n \right\} = \mathbb{P}_\pi\left\{ \left(\overline{Y}_{1,\pi} - \overline{Y}_{2,\pi}\right) \geq \frac{tn^2}{2n_1n_2} \middle| \mathcal{X}_n \right\} \leq \exp\left\{ -\frac{n^2t^2}{4\sum_{i=1}^n (Y_i - \overline{Y})^2} \right\},$$

where $\overline{Y}_{1,\pi} = n_1^{-1}\sum_{i=1}^n Y_{\pi_i}\mathbb{1}(Z_i = 1)$ and $\overline{Y}_{2,\pi} = n_2^{-1}\sum_{i=1}^n Y_{\pi_i}\mathbb{1}(Z_i = -1)$. This implies that

$$\mathbb{P}_\pi\left(\overline{Y}_{1,\pi} - \overline{Y}_{2,\pi} \geq t \middle| \mathcal{X}_n\right) \leq \exp\left(-\frac{n_1^2n_2^2t^2}{n^3\widehat{\sigma}_{\text{lin}}^2}\right),$$

where $\widehat{\sigma}_{\text{lin}}^2 = n^{-1}\sum_{i=1}^n (Z_i - \overline{Z})^2$. By symmetry, it also holds that

$$\mathbb{P}_\pi\left(|\overline{Y}_{1,\pi} - \overline{Y}_{2,\pi}| \geq t \middle| \mathcal{X}_n\right) \leq 2\exp\left(-\frac{n_1^2n_2^2t^2}{n^3\widehat{\sigma}_{\text{lin}}^2}\right).$$

Let us denote the sample mean of the entire samples by $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$. Then using the exact relationship

$$|\overline{Y}_{1,\pi} - \overline{Y}_{2,\pi}| = \frac{n}{n_2}|\overline{Y}_{1,\pi} - \overline{Y}|, \tag{G.5}$$

the above inequality is equivalent to

$$\mathbb{P}_\pi\left(|\overline{Y}_{1,\pi} - \overline{Y}| \geq t \middle| \mathcal{X}_n\right) \leq 2\exp\left(-\frac{n_1^2t^2}{n\widehat{\sigma}_{\text{lin}}^2}\right). \tag{G.6}$$

Notice that $\overline{Y}_{1,\pi}$ is the sample mean of n_1 observations sampled without replacement from $\{Y_1, \ldots, Y_n\}$. This implies that the permutation law of the sample mean is equivalent to the probability law under sampling without replacement. The same result (including the constant factor) exists in Massart (1986) (see Lemma 3.1 therein). However, the result given there only holds when $n = n_1 \times m$ where m is a positive integer whereas our result does not require such restriction.

An improvement via Berstein-type bound. Although the tail bound (G.6) is simple depending only on the variance term $\widehat{\sigma}_{\text{lin}}^2$, it may not be effective when n_1 is much smaller than n (e.g. $n_1^2/n \to 0$ as $n_1 \to \infty$). In such case, Proposition G.2 gives a tighter bound. More specifically, following the same steps as before, Proposition G.2 presents a concentration inequality for the two-sample (unscaled) t-statistic as

$$\mathbb{P}_\pi(L_n^\pi \geq t|\mathcal{X}_n) = \mathbb{P}_\pi\left\{ \left(\overline{Y}_{1,\pi} - \overline{Y}_{2,\pi}\right) \geq \frac{tn^2}{2n_1n_2} \middle| \mathcal{X}_n \right\}$$

$$\leq \exp\left\{ - \frac{nt^2}{8\frac{n_1 n_2}{n^2}\widehat{\sigma}^2_{\mathrm{lin}} + \frac{4}{3}t \cdot \max\left(\frac{n_1}{n}, \frac{n_2}{n}\right) \cdot M_Z} \right\},$$

where $M_Z := \max_{1 \leq i \leq n} |Z_i - \overline{Z}|$. Furthermore, using the relationship (G.5) and by symmetry,

$$\mathbb{P}_\pi \left(|\overline{Y}_{1,\pi} - \overline{Y}| \geq t \middle| \mathcal{X}_n \right) \leq 2 \exp\left\{ - \frac{12 n_1 t^2}{24\frac{n_2}{n}\widehat{\sigma}^2_{\mathrm{lin}} + 8\frac{n_2}{n} M_Z t} \right\}, \tag{G.7}$$

where we assumed $n_1 \leq n_2$.

Remark G.1. We remark that the bounds in (G.6) and (G.7) are byproducts of more general bounds and are not necessary the sharpest ones in the context of sampling without replacement. We refer to Bardenet and Maillard (2015) and among others for some recent developments of concentration bounds for sampling without replacement.

G.3 Improved version of Theorem 8.1

In this section, we improve the result of Theorem 8.1 based on the exponential bound in Theorem 8.3. In particular we replace the dependency on α^{-1} there with $\log(1/\alpha)$ by adding an extra assumption that $n_1 \asymp n_2$ as follows.

Lemma G.0.1 (Two-sample U-statistic)**.** *For $0 < \alpha < e^{-1}$, suppose that there is a sufficiently large constant $C > 0$ such that*

$$\mathbb{E}_P[U_{n_1,n_2}] \geq C \max\left\{ \sqrt{\frac{\psi_{Y,1}(P)}{\beta n_1}}, \sqrt{\frac{\psi_{Z,1}(P)}{\beta n_2}}, \sqrt{\frac{\psi_{YZ,2}(P)}{\beta}} \log\left(\frac{1}{\alpha}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \right\}, \tag{G.8}$$

for all $P \in \mathcal{P}_1 \subset \mathcal{P}_{h_{\mathrm{ts}}}$. Then under the assumptions that $n_1 \asymp n_2$, the type II error of the permutation test over $\mathcal{P}_1$ is uniformly bounded by β, that is

$$\sup_{P \in \mathcal{P}_1} \mathbb{P}_P^{(n_1,n_2)}(U_{n_1,n_2} \leq c_{1-\alpha,n_1,n_2}) \leq \beta.$$

Proof. To prove the above lemma, we employ the quantile approach described in Section 8.3 (see also Fromont et al., 2013). More specifically we let $q_{1-\beta/2,n}$ denote the quantile of the permutation critical value $c_{1-\alpha,n}$ of U_{n_1,n_2}. Then as shown in the proof of Lemma 8.0.1, if

$$\mathbb{E}_P[U_{n_1,n_2}] \;\geq\; q_{1-\beta/2,n} + \sqrt{\frac{2\mathrm{Var}_P[U_{n_1,n_2}]}{\beta}},$$

then the type II error of the permutation test is controlled as

$$\begin{aligned}\sup_{P\in\mathcal{P}_1}\mathbb{P}_P(U_{n_1,n_2}\le c_{1-\alpha,n}) &\le \sup_{P\in\mathcal{P}_1}\mathbb{P}_P(U_{n_1,n_2}\le q_{1-\beta/2,n})+\sup_{P\in\mathcal{P}_1}\mathbb{P}_P(q_{1-\beta/2,n}<c_{1-\alpha,n})\\ &\le \beta.\end{aligned}$$

Therefore it is enough to verify that the right-hand side of (G.8) is lower bounded by $q_{1-\beta/2,n}+\sqrt{2\mathrm{Var}_P[U_{n_1,n_2}]/\beta}$. As shown in the proof of Theorem 8.1, the variance is bounded by

$$\mathrm{Var}_P[U_{n_1,n_2}]\le C_1\frac{\psi_{Y,1}(P)}{n_1}+C_2\frac{\psi_{Z,1}(P)}{n_2}+C_3\psi_{YZ,2}(P)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^2. \tag{G.9}$$

Moving onto an upper bound for $q_{1-\beta/2,n}$, let us denote

$$\Sigma^{\dagger}_{n_1,n_2}:=\frac{1}{n_1^2(n_1-1)^2}\sum_{(i_1,i_2)\in\mathbf{i}_2^n}g^2(X_{i_1},X_{i_2}).$$

From Theorem 8.3 together with the trivial bound (8.31), we know that $c_{1-\alpha,n}$ is bounded by

$$\begin{aligned}c_{1-\alpha,n} &\le \max\left\{\sqrt{\frac{\Sigma^{\dagger 2}_{n_1,n_2}}{C_4}\log\left(\frac{1}{\alpha}\right)},\ \frac{\Sigma^{\dagger}_{n_1,n_2}}{C_4}\log\left(\frac{1}{\alpha}\right)\right\}\\ &\le C_5\Sigma^{\dagger}_{n_1,n_2}\log\left(\frac{1}{\alpha}\right),\end{aligned} \tag{G.10}$$

where the last inequality uses the assumption that $\alpha<e^{-1}$. Now applying Markov's inequality yields

$$\mathbb{P}_P\left(\Sigma^{\dagger}_{n_1,n_2}\ge t\right)\le\frac{\mathbb{E}_P[\Sigma^{\dagger 2}_{n_1,n_2}]}{t^2}\le C_6\frac{\psi_{YZ,2}(P)}{t^2n_1^2}.$$

By setting the right-hand side to be $\beta/2$, we can find an upper bound for the $1-\beta/2$ quantile of $\Sigma^{\dagger}_{n_1,n_2}$. Combining this observation with inequality (G.10) yields

$$q_{1-\beta/2,n}\le\frac{C_7}{\beta^{1/2}}\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{\psi_{YZ,2}(P)}}{n_1}.$$

Therefore, from the above bound and (G.9),

$$\begin{aligned}&q_{1-\beta/2,n}+\sqrt{\frac{2\mathrm{Var}_P[U_{n_1,n_2}]}{\beta}}\\ &\le C\sqrt{\max\left\{\frac{\psi_{Y,1}(P)}{\beta n_1},\ \frac{\psi_{Z,1}(P)}{\beta n_2},\ \frac{\psi_{YZ,2}(P)}{\beta}\log^2\left(\frac{1}{\alpha}\right)\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^2\right\}}.\end{aligned}$$

This completes the proof of Lemma G.0.1. □

G.4 Proof of Lemma 8.0.1

As discussed in the main text, the key difficulty of studying the type II error of the permutation test lies in the fact that its critical value is data-dependent and thereby random. Our strategy to overcome this problem is to bound the random critical value by a quantile value with high probability (see also Fromont et al., 2013). We split the proof of Lemma 8.0.1 into three steps. In the first step, we present a sufficient condition under which the type II error of the test with a non-random cutoff value is small. In the second step, we provide a non-random upper bound for the permutation critical value, which holds with high probability. In the last step, we combine the results and complete the proof.

Step 1. For a given $P \in \mathcal{P}_1$, let $\omega(P)$ be any constant depending on P such that

$$\mathbb{E}_P[T_n] \;\geq\; \omega(P) + \sqrt{\frac{3\mathrm{Var}_P[T_n]}{\beta}}. \tag{G.11}$$

Based on such $\omega(P)$, we define a test $\mathbb{1}\{T_n > \omega(P)\}$, which controls the type II error by $\beta/3$. To see this, let us apply Chebyshev's inequality

$$\begin{aligned}
\beta/3 \;&\geq\; \mathbb{P}_P\big(\big|T_n - \mathbb{E}_P[T_n]\big| \geq \sqrt{3\beta^{-1}\mathrm{Var}_P[T_n]}\big) \\
&\geq\; \mathbb{P}_P\big(-T_n + \mathbb{E}_P[T_n] \geq \sqrt{3\beta^{-1}\mathrm{Var}_P[T_n]}\big) \\
&\geq\; \mathbb{P}_P\big(\omega(P) \geq T_n\big),
\end{aligned}$$

where the last inequality uses the condition of $\omega(P)$ in (G.11). In other words, the type II error of the test $\mathbb{1}\{T_n > \omega(P)\}$ is less than or equal to $\beta/3$ as desired.

Step 2. In this step, we provide an upper bound for $c_{1-\alpha,n}$, which may hold with high probability. First, applying Chebyshev's inequality yields

$$\mathbb{P}_\pi\left(\big|T_n^\pi - \mathbb{E}_\pi[T_n^\pi|\mathcal{X}_n]\big| \geq \sqrt{\alpha^{-1}\mathrm{Var}_\pi[T_n^\pi|\mathcal{X}_n]}\;\big|\mathcal{X}_n\right) \leq \alpha.$$

Therefore, by the definition of the quantile, we see that $c_{1-\alpha,n}$ satisfies

$$c_{1-\alpha,n} \;\leq\; \mathbb{E}_\pi[T_n^\pi|\mathcal{X}_n] + \sqrt{\alpha^{-1}\mathrm{Var}_\pi[T_n^\pi|\mathcal{X}_n]}. \tag{G.12}$$

Note that the two terms on the right-hand side are random variables depending on $\mathcal{X}_n$. In order to use the result from the first step, we want to further upper bound these two terms by some constants. To this end, let us define two good events:

$$\mathcal{A}_1 := \Big\{ \mathbb{E}_\pi[T_n^\pi | \mathcal{X}_n] < \mathbb{E}_P[\mathbb{E}_\pi\{T_n^\pi | \mathcal{X}_n\}] + \sqrt{3\beta^{-1}\mathrm{Var}_P[\mathbb{E}_\pi\{T_n^\pi | \mathcal{X}_n\}]} \Big\},$$
$$\mathcal{A}_2 := \Big\{ \sqrt{\alpha^{-1}\mathrm{Var}_\pi[T_n^\pi | \mathcal{X}_n]} < \sqrt{3\alpha^{-1}\beta^{-1}\mathbb{E}_P[\mathrm{Var}_\pi\{T_n^\pi | \mathcal{X}_n\}]} \Big\}.$$

Then by applying Markov and Chebyshev's inequalities, it is straightforward to see that

$$\mathbb{P}_P(\mathcal{A}_1^c) \le \beta/3 \quad \text{and} \quad \mathbb{P}_P(\mathcal{A}_2^c) \le \beta/3. \tag{G.13}$$

Step 3. Here, building on the first two steps, we conclude the result. We begin by upper bounding the type II error of the permutation test as

$$\begin{aligned} \mathbb{P}_P(T_n \le c_{1-\alpha,n}) &= \mathbb{P}_P(T_n \le c_{1-\alpha,n},\ \mathcal{A}_1 \cup \mathcal{A}_2) + \mathbb{P}_P(T_n \le c_{1-\alpha,n},\ \mathcal{A}_1^c \cap \mathcal{A}_2^c) \\ &\le \mathbb{P}_P(T_n \le \omega'(P)) + \mathbb{P}_P(\mathcal{A}_1^c \cap \mathcal{A}_2^c), \end{aligned}$$

where, for simplicity, we write

$$\omega'(P) := \mathbb{E}_P[\mathbb{E}_\pi\{T_n^\pi | \mathcal{X}_n\}] + \sqrt{3\beta^{-1}\mathrm{Var}_P[\mathbb{E}_\pi\{T_n^\pi | \mathcal{X}_n\}]} + \sqrt{3\alpha^{-1}\beta^{-1}\mathbb{E}_P[\mathrm{Var}_\pi\{T_n^\pi | \mathcal{X}_n\}]}.$$

One may check that the type II error of $\mathbb{1}\{T_n > \omega'(P)\}$ is controlled by $\beta/3$ as long as $\omega'(P) + \sqrt{3\mathrm{Var}_P[T_n]/\beta} \le \mathbb{E}_P[T_n]$ from the inequality (G.11) in Step 1. However, this sufficient condition is ensured by condition (8.3) of Lemma 8.0.1. Furthermore, the probability of the intersection of the two bad events $\mathcal{A}_1^c \cap \mathcal{A}_2^c$ is also bounded by $2\beta/3$ due to the concentration results in (G.13). Hence, by taking the supremum over $P \in \mathcal{P}_1$, we may conclude that

$$\sup_{P \in \mathcal{P}_1} \mathbb{P}_P(T_n \le c_{1-\alpha,n}) \le \beta.$$

This completes the proof of Lemma 8.0.1.

G.5 Proof of Theorem 8.1

We proceed the proof by verifying the sufficient condition in Lemma 8.0.1. We first verify that the expectation of U_{n_1,n_2}^π is zero under the permutation law. Let us recall the permuted U-statistic U_{n_1,n_2}^π in (8.26). In fact,

by the linearity of expectation, it suffices to prove

$$\mathbb{E}_\pi[h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_2}; X_{\pi_{n_1+1}}, X_{\pi_{n_1+2}})|\mathcal{X}_n] = 0.$$

This is clearly the case by recalling the definition of kernel h_{ts} in (8.5) and noting that the expectation $\mathbb{E}_\pi[g(X_{\pi_i}, X_{\pi_j})|\mathcal{X}_n]$ is invariant to the choice of $(i,j) \in \mathbf{i}_2^n$, which leads to $\mathbb{E}_\pi[U^\pi_{n_1,n_2}|\mathcal{X}_n] = 0$. Therefore we only need to verify the simplified condition (8.4) under the given assumptions in Theorem 8.1.

The rest of the proof is divided into two parts. In each part, we prove the following conditions separately,

$$\mathbb{E}_P[U_{n_1,n_2}] \geq 2\sqrt{\frac{2\mathrm{Var}_P[U_{n_1,n_2}]}{\beta}} \quad \text{and} \tag{G.14}$$

$$\mathbb{E}_P[U_{n_1,n_2}] \geq 2\sqrt{\frac{2\mathbb{E}_P[\mathrm{Var}_\pi\{U^\pi_{n_1,n_2}|\mathcal{X}_n\}]}{\alpha\beta}}. \tag{G.15}$$

We then complete the proof of Theorem 8.1 by noting that (G.14) and (G.15) imply the simplified condition (8.4).

Part 1. Verification of condition (G.14): In this part, we verify condition (G.14). To do so, we state the explicit variance formula of a two-sample U-statistic (e.g. page 38 of Lee, 1990). Following the notation of Lee (1990), we let $\check{\sigma}^2_{i,j}$ denote the variance of a conditional expectation given as

$$\check{\sigma}^2_{i,j} = \mathrm{Var}_P[\mathbb{E}_P\{\overline{h}_{\mathrm{ts}}(y_1, \ldots, y_i, Y_{i+1}, \ldots, Y_2; z_1, \ldots, z_j, Z_{j+1}, \ldots, Z_2)\}] \quad \text{for } 0 \leq i,j \leq 2.$$

Then the variance of U_{n_1,n_2} is given by

$$\mathrm{Var}_P[U_{n_1,n_2}] = \sum_{i=0}^{2}\sum_{j=0}^{2} \binom{2}{i}\binom{2}{j}\binom{n_1-2}{2-i}\binom{n_2-2}{2-j}\binom{n_1}{2}^{-1}\binom{n_2}{2}^{-1}\check{\sigma}^2_{i,j}. \tag{G.16}$$

By the law of total variance, one may see that $\check{\sigma}^2_{i,j} \leq \check{\sigma}^2_{2,2}$ for all $0 \leq i,j \leq 2$. This leads to an upper bound for $\mathrm{Var}_P[U_{n_1,n_2}]$ as

$$\mathrm{Var}_P[U_{n_1,n_2}] \leq C_1\frac{\check{\sigma}^2_{1,0}}{n_1} + C_2\frac{\check{\sigma}^2_{0,1}}{n_2} + C_3\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^2\check{\sigma}^2_{2,2}.$$

Now applying Jensen's inequality, repeatedly, yields

$$\check{\sigma}^2_{2,2} \leq \mathbb{E}_P[\overline{h}^2_{\mathrm{ts}}(Y_1, Y_2; Z_1, Z_2)] \leq \mathbb{E}_P[h^2_{\mathrm{ts}}(Y_1, Y_2; Z_1, Z_2)] \leq C_4\psi_{YZ,2}(P).$$

Then by noting that $\check{\sigma}^2_{1,0}$ and $\check{\sigma}^2_{0,1}$ correspond to the notation $\psi_{Y,1}(P)$ and $\psi_{Z,1}(P)$, respectively,

$$\mathrm{Var}_P[U_{n_1,n_2}] \leq C_1 \frac{\psi_{Y,1}(P)}{n_1} + C_2 \frac{\psi_{Z,1}(P)}{n_2} + C_4 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2 \psi_{YZ,2}(P).$$

Hence condition (G.14) is satisfied by taking the constant C in Theorem 8.1 sufficiently large.

Part 2. Verification of condition (G.15): In this part, we verify condition (G.15). Intuitively, the permuted U-statistic behaves similarly as the unconditional U-statistic under a certain null model. This means that the variance of $U^{\pi}_{n_1,n_2}$ should have a similar convergence rate as $(n_1^{-1} + n_2^{-1})^2 \psi_{YZ,2}(P)$ since $\psi_{Y,1}(P)$ and $\psi_{Z,1}(P)$ are zero under the null hypothesis. We now prove that this intuition is indeed correct. Since $U^{\pi}_{n_1,n_2}$ is centered under the permutation law, it is enough to study $\mathbb{E}_P[\mathbb{E}_\pi\{(U^{\pi}_{n_1,n_2})^2 | \mathcal{X}_n\}]$. Let us write a set of indices $\mathsf{I}_{\mathrm{total}} := \{(i_1, i_2, j_1, j_2, i'_1, i'_2, j'_1, j'_2) \in \mathbb{N}^8_+ : (i_1, i_2) \in \mathbf{i}_2^{n_1}, (j_1, j_2) \in \mathbf{i}_2^{n_2}, (i'_1, i'_2) \in \mathbf{i}_2^{n_1}, (j'_1, j'_2) \in \mathbf{i}_2^{n_2}\}$ and define $\mathsf{I}_{\mathrm{A}} = \{(i_1, i_2, j_1, j_2, i'_1, i'_2, j'_1, j'_2) \in \mathsf{I}_{\mathrm{total}} : \#|\{i_1, i_2, j_1, j_2\} \cap \{i'_1, i'_2, j'_1, j_2\}| \leq 1\}$ and $\mathsf{I}_{\mathrm{A}^c} = \{(i_1, i_2, j_1, j_2, i'_1, i'_2, j'_1, j'_2) \in \mathsf{I}_{\mathrm{total}} : \#|\{i_1, i_2, j_1, j_2\} \cap \{i'_1, i'_2, j'_1, j_2\}| > 1\}$. Here $\#|B|$ denotes the cardinality of a set B. Then it is clear that $\mathsf{I}_{\mathrm{total}} = \mathsf{I}_{\mathrm{A}} \cup \mathsf{I}_{\mathrm{A}^c}$. Based on this notation and the linearity of expectation,

$$\begin{aligned}
\mathbb{E}_\pi[(U^{\pi}_{n_1,n_2})^2 | \mathcal{X}_n] &= \frac{1}{(n_1)^2_{(2)}(n_2)^2_{(2)}} \sum_{(i_1,\ldots,j'_2) \in \mathsf{I}_{\mathrm{total}}} \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_{i_1}}, X_{\pi_{i_2}}; X_{\pi_{n_1+j_1}}, X_{\pi_{n_1+j_2}}) \\
&\qquad \times h_{\mathrm{ts}}(X_{\pi_{i'_1}}, X_{\pi_{i'_2}}; X_{\pi_{n_1+j'_1}}, X_{\pi_{n_1+j'_2}}) \,\Big|\, \mathcal{X}_n \Big] \\
&= (I) + (II),
\end{aligned}$$

where

$$\begin{aligned}
(I) &:= \frac{1}{(n_1)^2_{(2)}(n_2)^2_{(2)}} \sum_{(i_1,\ldots,j'_2) \in \mathsf{I}_{\mathrm{A}}} \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_{i_1}}, X_{\pi_{i_2}}; X_{\pi_{n_1+j_1}}, X_{\pi_{n_1+j_2}}) \\
&\qquad \times h_{\mathrm{ts}}(X_{\pi_{i'_1}}, X_{\pi_{i'_2}}; X_{\pi_{n_1+j'_1}}, X_{\pi_{n_1+j'_2}}) \,\Big|\, \mathcal{X}_n \Big], \\
(II) &:= \frac{1}{(n_1)^2_{(2)}(n_2)^2_{(2)}} \sum_{(i_1,\ldots,j'_2) \in \mathsf{I}_{\mathrm{A}^c}} \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_{i_1}}, X_{\pi_{i_2}}; X_{\pi_{n_1+j_1}}, X_{\pi_{n_1+j_2}}) \\
&\qquad \times h_{\mathrm{ts}}(X_{\pi_{i'_1}}, X_{\pi_{i'_2}}; X_{\pi_{n_1+j'_1}}, X_{\pi_{n_1+j'_2}}) \,\Big|\, \mathcal{X}_n \Big].
\end{aligned}$$

We now claim that the first term $(I) = 0$. This is the key observation that makes the upper bound for the variance of the permuted U-statistic depend on $(n_1^{-1} + n_2^{-1})^2$ rather than a slower rate $(n_1 + n_2)^{-1}$. First consider the case where $\#|\{i_1, i_2, j_1, j_2\} \cap \{i'_1, i'_2, j'_1, j_2\}| = 0$, that is, all indices are distinct. Let us focus on the summands of (I). By symmetry, we may assume the set of indices $(i_1, i_2, n_1 + j_1, n_1 + j_2, i'_1, i'_2, n_1 +$

$j_1', n_1 + j_2')$ to be $(1, \ldots, 8)$ and observe that

$$\begin{aligned}
& \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_2}; X_{\pi_3}, X_{\pi_4}) h_{\mathrm{ts}}(X_{\pi_5}, X_{\pi_6}; X_{\pi_7}, X_{\pi_8}) \Big| \mathcal{X}_n \Big] \\
\overset{(i)_1}{=} \; & \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_3}, X_{\pi_2}; X_{\pi_1}, X_{\pi_4}) h_{\mathrm{ts}}(X_{\pi_5}, X_{\pi_6}; X_{\pi_7}, X_{\pi_8}) \Big| \mathcal{X}_n \Big] \\
\overset{(ii)_1}{=} \; & - \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_2}; X_{\pi_3}, X_{\pi_4}) h_{\mathrm{ts}}(X_{\pi_5}, X_{\pi_6}; X_{\pi_7}, X_{\pi_8}) \Big| \mathcal{X}_n \Big] \\
\overset{(iii)_1}{=} \; & 0,
\end{aligned}$$

where $(i)_1$ holds since the distribution of the product kernels does not change even after π_1 and π_3 are switched and $(ii)_1$ uses the fact that $h_{\mathrm{ts}}(y_1, y_2; z_1, z_2) = -h_{\mathrm{ts}}(z_1, y_2; y_1, z_2)$. $(iii)_1$ follows directly by comparing the first line and the third line of the equations. Next consider the case where $\#|\{i_1, i_2, j_1, j_2\} \cap \{i_1', i_2', j_1', j_2\}| = 1$. Without loss of generality, assume that $i_1 = i_1'$. In this case, by symmetry again, we have

$$\begin{aligned}
& \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_2}; X_{\pi_3}, X_{\pi_4}) h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_5}; X_{\pi_6}, X_{\pi_7}) \Big| \mathcal{X}_n \Big] \\
\overset{(i)_2}{=} \; & \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_4}; X_{\pi_3}, X_{\pi_2}) h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_5}; X_{\pi_6}, X_{\pi_7}) \Big| \mathcal{X}_n \Big] \\
\overset{(ii)_2}{=} \; & - \mathbb{E}_\pi \Big[h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_2}; X_{\pi_3}, X_{\pi_4}) h_{\mathrm{ts}}(X_{\pi_1}, X_{\pi_5}; X_{\pi_6}, X_{\pi_7}) \Big| \mathcal{X}_n \Big] \\
\overset{(iii)_2}{=} \; & 0,
\end{aligned}$$

where $(i)_2$ follows by the same reasoning for $(i)_2$ and $(ii)_2$ holds since $h_{\mathrm{ts}}(y_1, y_2; z_1, z_2) = -h_{\mathrm{ts}}(y_1, z_2; y_1, y_2)$. Then $(iii)_2$ is obvious by comparing the first line and the third line of the equations. Hence, for any choice of indices $(i_1, \ldots, j_2') \in \mathsf{I}_{\mathrm{A}}$, the summands of (I) becomes zero, which leads to $(I) = 0$.

Now turning to the second term (II), for any $1 \leq i_1 \neq i_2, i_3 \neq i_4 \leq n$, we have

$$\begin{aligned}
& \big| \mathbb{E}_P \big[\mathbb{E}_\pi \{ g(X_{\pi_{i_1}}, X_{\pi_{i_2}}) g(X_{\pi_{i_3}}, X_{\pi_{i_4}}) | \mathcal{X}_n \} \big] \big| \\
\overset{(i)_3}{=} \; & \big| \mathbb{E}_\pi \big[\mathbb{E}_P \{ g(X_{\pi_{i_1}}, X_{\pi_{i_2}}) g(X_{\pi_{i_3}}, X_{\pi_{i_4}}) | \pi_{i_1}, \ldots, \pi_{i_4} \} \big] \big| \\
\overset{(ii)_3}{\leq} \; & \frac{1}{2} \mathbb{E}_\pi \big[\mathbb{E}_P \{ g^2(X_{\pi_{i_1}}, X_{\pi_{i_2}}) | \pi_{i_1}, \pi_{i_2} \} \big] + \frac{1}{2} \mathbb{E}_\pi \big[\mathbb{E}_P \{ g^2(X_{\pi_{i_3}}, X_{\pi_{i_4}}) | \pi_{i_3}, \pi_{i_4} \} \big] \\
\overset{(iii)_3}{\leq} \; & \psi_{YZ,2}(P),
\end{aligned}$$

where $(i)_3$ uses the the law of total expectation, $(ii)_3$ uses the basic inequality $xy \leq x^2/2 + y^2/2$ and $(iii)_3$ clearly holds by recalling the definition of $\psi_{YZ,2}(P)$. Using this observation, it is not difficult to see that for

any $(i_1, \ldots, j_2') \in \mathsf{I}_{\text{total}}$,

$$\left|\mathbb{E}_\pi\left[h_{\text{ts}}(X_{\pi_{i_1}}, X_{\pi_{i_2}}; X_{\pi_{n_1+j_1}}, X_{\pi_{n_1+j_2}}) h_{\text{ts}}(X_{\pi_{i_1'}}, X_{\pi_{i_2'}}; X_{\pi_{n_1+j_1'}}, X_{\pi_{n_1+j_2'}}) \middle| \mathcal{X}_n\right]\right| \leq C_5 \psi_{YZ,2}(P).$$

Therefore, by counting the number of elements in $\mathsf{I}_{\mathrm{A}^c}$,

$$\begin{aligned}
\mathbb{E}_P[\mathrm{Var}_\pi\{U^\pi_{n_1,n_2} | \mathcal{X}_n\}] = \mathbb{E}_P[(II)] &\leq C_5 \psi_{YZ,2}(P) \times \frac{1}{(n_1)^2_{(2)} (n_2)^2_{(2)}} \sum_{(i_1,\ldots,j_2') \in \mathsf{I}_{\mathrm{A}^c}} 1 \\
&\leq C_6 \psi_{YZ,2}(P) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2.
\end{aligned}$$

Hence condition (G.15) is satisfied by taking the constant C in Theorem 8.1 sufficiently large. This completes the proof of Theorem 8.1.

G.6 Proof of Proposition 8.1

As discussed in the main text, we start proving that the three inequalities in (8.14) are fulfilled. Focusing on the first one, we want to show that

$$\psi_{Y,1}(P) \leq C_1 \sqrt{b_{(1)}} \|p_Y - p_Z\|_2^2 \quad \text{for some } C_1 > 0.$$

By denoting the kth component of p_Y and p_Z by $p_Y(k)$ and $p_Z(k)$, respectively, note that

$$\mathbb{E}_P[\overline{h}_{\text{ts}}(Y_1, Y_2; Z_1, Z_2) | Y_1] = \sum_{k=1}^d [\mathbb{1}(Y_1 = k) - p_Z(k)][p_Y(k) - p_Z(k)]$$

and so $\psi_{Y,1}(P)$, which is the variance of the above expression, becomes

$$\psi_{Y,1}(P) = \mathbb{E}_P\left[\left(\sum_{k=1}^d [\mathbb{1}(Y_1 = k) - p_Y(k)][p_Y(k) - p_Z(k)]\right)^2\right].$$

Furthermore, observe that

$$\begin{aligned}
\psi_{Y,1}(P) &\overset{(i)}{\leq} 2\mathbb{E}_P\left[\left(\sum_{k=1}^d \mathbb{1}(Y_1 = k)[p_Y(k) - p_Z(k)]\right)^2\right] + 2\left(\sum_{k=1}^d p_Y(k)[p_Y(k) - p_Z(k)]\right)^2 \\
&= 2\sum_{k=1}^d p_Y(k)[p_Y(k) - p_Z(k)]^2 + 2\left(\sum_{k=1}^d p_Y(k)[p_Y(k) - p_Z(k)]\right)^2
\end{aligned}$$

$$\overset{(ii)}{\leq} 2\sqrt{\sum_{k=1}^{d} p_Y^2(k)}\sqrt{\sum_{k=1}^{d}[p_Y(k)-p_Z(k)]^4} + 2\sum_{k=1}^{d} p_Y^2(k)\sum_{k=1}^{d}[p_Y(k)-p_Z(k)]^2$$

$$\overset{(iii)}{\leq} 4\sqrt{b_{(1)}}\|p_Y - p_Z\|_2^2,$$

where (i) is based on $(x+y)^2 \leq 2x^2+2y^2$, (ii) uses Cauchy-Schwarz inequality and (iii) uses the monotonicity of ℓ_p norm (specifically, $\ell_4 \leq \ell_2$) as well as the fact that $\|p_Y\|_2^2 \leq \|p_Y\|_2$. By symmetry, we can also have that

$$\psi_{Z,1}(P) \;\leq\; 4\sqrt{b_{(1)}}\|p_Y - p_Z\|_2^2.$$

Now focusing on the third line of the claim (8.14), recall that

$$\psi_{YZ,2}(P) := \max\{\mathbb{E}_P[g^2_{\text{Multi}}(Y_1,Y_2)],\ \mathbb{E}_P[g^2_{\text{Multi}}(Y_1,Z_1)],\ \mathbb{E}_P[g^2_{\text{Multi}}(Z_1,Z_2)]\}$$

and by noting that $g_{\text{Multi}}(x,y)$ is either one or zero,

$$\mathbb{E}_P[g^2_{\text{Multi}}(Y_1,Y_2)] = \sum_{k=1}^{d} p_Y^2(k),$$

$$\mathbb{E}_P[g^2_{\text{Multi}}(Z_1,Z_2)] = \sum_{k=1}^{d} p_Z^2(k) \quad \text{and}$$

$$\mathbb{E}_P[g^2_{\text{Multi}}(Y_1,Z_1)] = \sum_{k=1}^{d} p_Y(k)p_Z(k) \leq \frac{1}{2}\sum_{k=1}^{d} p_Y^2(k) + \frac{1}{2}\sum_{k=1}^{d} p_Z^2(k),$$

where the last inequality uses $xy \leq x^2/2 + y^2/2$. This clearly shows that $\psi_{YZ,2} \leq b_{(1)}$, which confirms the claim (8.14). Since the expectation of U_{n_1,n_2} is $\|p_Y - p_Z\|_2^2$, one may see that

$$\mathbb{E}_P[U_{n_1,n_2}] \geq \epsilon^2_{n_1,n_2} \geq C_1\frac{\sqrt{b_{(1)}}}{\alpha^{1/2}\beta}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

$$\geq C_2\sqrt{\max\left\{\frac{\psi_{Y,1}(P)}{\beta n_1},\ \frac{\psi_{Z,1}(P)}{\beta n_2},\ \frac{\psi_{YZ,2}(P)}{\alpha\beta}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^2\right\}}.$$

Now we apply Theorem 8.1 and finish the proof of Proposition 8.1.

G.7 Proof of Proposition 8.2

We first note that Proposition 8.1 establishes an upper bound for the minimum separation as $\epsilon^{\dagger}_{n_1,n_2} \lesssim b_{(1)}^{1/4} n_1^{-1/2}$ where $n_1 \leq n_2$. Hence once we identify a lower bound such that $\epsilon^{\dagger}_{n_1,n_2} \gtrsim b_{(1)}^{1/4} n_1^{-1/2}$, the proof is completed. As briefly explained in the main text, our strategy to prove this result is to consider the one-sample problem, which is conceptually easier than the two-sample problem, and establish the matching lower bound. In the one-sample problem, we assume that p_Z is known and observe n_1 samples from the other distribution p_Y. Based on these n_1 samples, we want to test whether $p_Y = p_Z$ or $\|p_Y - p_Z\|_2 \geq \epsilon_{n_1}$. As formalized by Arias-Castro et al. (2018) (see their Lemma 1), the one-sample problem can be viewed as a special case of the two-sample problem where one of the sample sizes is taken to be infinite and thus the minimum separation for the one-sample problem is always smaller than or equal to that for the two-sample problem. This means that if the minimum separation for the one-sample problem, denoted by $\epsilon^{\dagger}_{n_1}$, satisfies $\epsilon^{\dagger}_{n_1} \gtrsim b_{(1)}^{1/4} n_1^{-1/2}$, then we also have that $\epsilon^{\dagger}_{n_1,n_2} \gtrsim b_{(1)}^{1/4} n_1^{-1/2}$. In the end, it suffices to verify $\epsilon^{\dagger}_{n_1} \gtrsim b_{(1)}^{1/4} n_1^{-1/2}$ to complete the proof. We show this result based on the standard lower bound technique due to Ingster (1987, 1993).

• **Ingster's method for the lower bound.** Let us recall from Section 8.2.2 that the minimax type II error is given by

$$R^{\dagger}_{n,\epsilon_n} := \inf_{\phi \in \Phi_{n,\alpha}} \sup_{P \in \mathcal{P}_1(\epsilon_n)} \mathbb{P}_P^{(n)}(\phi = 0).$$

For $P_1, \ldots, P_N \in \mathcal{P}_1(\epsilon_n)$, define a mixture distribution Q given by

$$Q(A) = \frac{1}{N} \sum_{i=1}^{N} P_i^n(A).$$

Given n i.i.d. observations $X_1, \ldots, X_n$, we denote the likelihood ratio between Q and the null distribution P_0 by

$$L_n = \frac{dQ}{dP_0^n} = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{n} \frac{p_i(X_j)}{p_0(X_j)}.$$

Then one can relate the variance of the likelihood ratio to the minimax type II error as follows.

Lemma G.0.2 (Lower bound). *Let* $0 < \beta < 1 - \alpha$. *If*

$$\mathbb{E}_{P_0}[L_n^2] \leq 1 + 4(1 - \alpha - \beta)^2,$$

then $R^{\dagger}_{n,\epsilon_n} \geq \beta$.

Proof. We present the proof of this result only for completeness. Note that $\mathbb{P}^n_{P_0}(\phi = 1) \leq \alpha$ for $\phi \in \Phi_{n,\alpha}$. Thus

$$\begin{aligned} R^{\dagger}_{n,\epsilon_n} &\geq \inf_{\phi \in \Phi_{n,\alpha}} \mathbb{P}_Q(\phi = 0) = \inf_{\phi \in \Phi_{n,\alpha}} \big[\mathbb{P}_{P_0}(\phi = 0) + \mathbb{P}_Q(\phi = 0) - \mathbb{P}_{P_0}(\phi = 0)\big] \\ &\overset{(i)}{\geq} 1 - \alpha + \inf_{\phi \in \Phi_{n,\alpha}} \big[\mathbb{P}_Q(\phi = 0) - \mathbb{P}_{P_0}(\phi = 0)\big] \\ &\overset{(ii)}{\geq} 1 - \alpha - \sup_A \big|\mathbb{P}_Q(A) - \mathbb{P}_{P_0}(A)\big| \\ &\overset{(iii)}{=} 1 - \alpha - \frac{1}{2}\big\|Q - P_0^n\big\|_1. \end{aligned}$$

where (i) uses the fact that $\mathbb{P}^n_{P_0}(\phi = 1) \leq \alpha$, (ii) follows by taking the supremum over all measurable sets, (iii) uses the alternative expression for the total variation distance in terms of L_1-distance. The result then follows by noting that

$$\big\|Q - P_0^n\big\|_1 = \mathbb{E}_{P_0}[|L_n(X_1, \ldots, X_n) - 1|] \leq \sqrt{\mathbb{E}_{P_0}[L_n^2(X_1, \ldots, X_n)] - 1}.$$

This proves Lemma G.0.2. □

Next we apply this method to find a lower bound for $\epsilon^{\dagger}_{n_1}$. To apply Lemma G.0.2, we first construct Q and P_0.

• **Construction of Q and P_0.** Suppose that p_Z is the uniform distribution over $\mathbb{S}_d$, that is $p_Z(k) = 1/d$ for $k = 1, \ldots, d$. Let $\widetilde{\zeta} = \{\widetilde{\zeta}_1, \ldots, \widetilde{\zeta}_d\}$ be dependent Rademacher random variables uniformly distributed over $\{-1, 1\}^d$ such that $\sum_{i=1}^d \widetilde{\zeta}_i = 0$ where d is assumed to be even. More formally we define such a set by

$$\mathcal{M}_d := \big\{x \in \{-1, 1\}^d : \sum_{i=1}^d x_i = 0\big\}. \tag{G.17}$$

If d is odd, then we set $\widetilde{\zeta}_d = 0$ and the proof follows similarly. Given $\widetilde{\zeta} \in \mathcal{M}_d$, let us define a distribution $p_{\widetilde{\zeta}}$ as

$$p_{\widetilde{\zeta}}(k) := p_Z(k) + \delta \sum_{i=1}^d \widetilde{\zeta}_i \mathbb{1}(k = i),$$

where δ is specified later but $\delta \leq 1/d$. There are N such distributions where N is the cardinality of $\mathcal{M}_d$ and we denote them by $p_{\widetilde{\zeta}(1)}, \ldots, p_{\widetilde{\zeta}(N)}$. By construction we make three observations. First $p_{\widetilde{\zeta}}$ is a proper distribution as each component $p_{\widetilde{\zeta}}(k)$ is non-negative and $\sum_{k=1}^d p_{\widetilde{\zeta}}(k) = 1$. Second the ℓ_2 distance between

$p_{\widetilde{\zeta}}$ and p_Z is

$$\|p_{\widetilde{\zeta}} - p_Z\|_2 = \delta\sqrt{d}. \tag{G.18}$$

Third we see that $b_{(1)} = \max\{\|p_Z\|_2^2, \|p_{\widetilde{\zeta}}\|_2^2\}$ is lower and upper bounded by

$$\frac{1}{d} \leq b_{(1)} \leq \frac{2}{d}, \tag{G.19}$$

which can be verified based on Cauchy-Schwarz inequality and the fact that $\delta \leq 1/d$. Finally we denote the uniform mixture of $p_{\widetilde{\zeta}(1)}, \ldots, p_{\widetilde{\zeta}(N)}$ by

$$Q := \frac{1}{N}\sum_{i=1}^{N} p_{\widetilde{\zeta}(i)}$$

and let $P_0 = p_Z$. Having Q and P_0 at hand, we are now ready to compute the expected value of the squared likelihood ratio.

• **Calculation of $\mathbb{E}_{P_0}[L_n^2]$.** For each $\widetilde{\zeta}_{(i)} \in \mathcal{M}_d$ and $i = 1, \ldots, N$, let us denote the components of $\widetilde{\zeta}_{(i)}$ by $\{\widetilde{\zeta}_{1,(i)}, \ldots, \widetilde{\zeta}_{d,(i)}\}$. Based on this notation as well as the definition of Q and P_0, the squared the likelihood ratio L_n^2 can be written as

$$\begin{aligned}
L_n^2 &= \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\prod_{j=1}^{n_1}\frac{p_{\widetilde{\zeta}(i_1)}(X_j)p_{\widetilde{\zeta}(i_2)}(X_j)}{p_0(X_j)p_0(X_j)} \\
&= \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\prod_{j=1}^{n_1}\frac{\{1/d + \delta\sum_{k=1}^{d}\widetilde{\zeta}_{k(i_1)}\mathbb{1}(X_j = k)\}\{1/d + \delta\sum_{k=1}^{d}\widetilde{\zeta}_{k(i_2)}\mathbb{1}(X_j = k)\}}{1/d^2} \\
&= \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\prod_{j=1}^{n_1}\Big\{1 + d\delta\sum_{k=1}^{d}\widetilde{\zeta}_{k(i_1)}\mathbb{1}(X_j = k)\Big\}\Big\{1 + d\delta\sum_{k=1}^{d}\widetilde{\zeta}_{k(i_2)}\mathbb{1}(X_j = k)\Big\}.
\end{aligned}$$

Now by taking the expectation under P_0, it can be seen that

$$\begin{aligned}
\mathbb{E}_{P_0}[L_n^2] &= \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\left(1 + d\delta^2\sum_{k=1}^{d}\widetilde{\zeta}_{k(i_1)}\widetilde{\zeta}_{k(i_2)}\right)^{n_1} \\
&\leq \frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\exp\left(n_1 d\delta^2\sum_{k=1}^{d}\widetilde{\zeta}_{k(i_1)}\widetilde{\zeta}_{k(i_2)}\right),
\end{aligned}$$

where the inequality uses $1 + x \leq e^x$ for all $x \in \mathbb{R}$. By letting $\widetilde{\zeta}^*$ be i.i.d. copy of $\widetilde{\zeta}$, we may see that

$$\frac{1}{N^2}\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\exp\left(n_1 d\delta^2 \sum_{k=1}^{d}\widetilde{\zeta}_{k(i_1)}\widetilde{\zeta}_{k(i_2)}\right) = \mathbb{E}_{\widetilde{\zeta},\widetilde{\zeta}^*}\left[\exp\left(n_1 d\delta^2\langle\widetilde{\zeta},\widetilde{\zeta}^*\rangle\right)\right].$$

Moreover $\{\widetilde{\zeta}_1, \ldots \widetilde{\zeta}_d\}$ are negatively associated (Dubhashi and Ranjan, 1998). Hence applying Lemma 2 of Dubhashi and Ranjan (1998) yields

$$\begin{aligned}
\mathbb{E}_{P_0}[L_n^2] &\leq \mathbb{E}_{\widetilde{\zeta},\widetilde{\zeta}^*}\left[\exp\left(n_1 d\delta^2\langle\widetilde{\zeta},\widetilde{\zeta}^*\rangle\right)\right] \\
&\leq \prod_{i=1}^{d}\mathbb{E}_{\widetilde{\zeta}_i,\widetilde{\zeta}_i^*}\left[\exp\left(n_1 d\delta^2\widetilde{\zeta}_i\widetilde{\zeta}_i^*\right)\right] = \prod_{i=1}^{d}\cosh(n_1 d\delta^2) \\
&\overset{(i)}{\leq} \prod_{i=1}^{d} e^{n_1^2 d^2\delta^4/2} = e^{n_1^2 d^3\delta^4/2},
\end{aligned}$$

where (i) uses the inequality $\cosh(x) \leq e^{x^2/2}$ for all $x \in \mathbb{R}$.

• **Completion of the proof.** Based on this upper bound, we have from Lemma G.0.2 that if

$$\delta \leq \frac{1}{\sqrt{n_1}d^{3/4}}\left[\log\left\{1 + 4(1-\alpha-\beta)^2\right\}\right]^{1/4}$$

the minimax type II error is lower bounded by β. Furthermore, based on the expression for the ℓ_2 norm in (G.18) and the bound for $b_{(1)}$ in (G.19). The above condition is further implied by

$$\epsilon_{n_1} \leq \frac{b_{(1)}^{1/4}}{\sqrt{n_1}}\left[\log\left\{1 + 4(1-\alpha-\beta)^2\right\}\right]^{1/4}.$$

This completes the proof of Proposition 8.2.

G.8 Proof of Proposition 8.3

The proof of Proposition 8.3 is fairly straightforward based on Proposition 8.1 and Lemma 3 of Arias-Castro et al. (2018). For two vectors $\mathbf{v} = (v_1, \ldots, v_d) \in \mathbb{R}^d$ and $\mathbf{w} = (w_1, \ldots, w_d) \in \mathbb{R}^d$ where $v_i \leq w_i$ for all i, we borrow the notation from Arias-Castro et al. (2018) and denote the hyperrectangle by

$$[\mathbf{v}, \mathbf{w}] = \prod_{i=1}^{d}[v_i, w_i].$$

Recall that $\kappa_{(1)} = \lfloor n_1^{2/(4s+d)} \rfloor$ and define $\boldsymbol{H}_{\boldsymbol{\ell}} := [(\boldsymbol{\ell}-1)/\kappa_{(1)}, \boldsymbol{\ell}/\kappa_{(1)}]$ where $\boldsymbol{\ell} \in \{1, 2, \ldots, \kappa_{(1)}\}^d$,

$$p_Y(\boldsymbol{\ell}) := \int_{\boldsymbol{H}_{\boldsymbol{\ell}}} f_Y(t)dt \quad \text{and} \quad p_Z(\boldsymbol{\ell}) := \int_{\boldsymbol{H}_{\boldsymbol{\ell}}} f_Z(t)dt.$$

Since both f_Y and f_Z are in Hölder's density class $\mathcal{P}^{(d,s)}_{\text{Hölder}}$ where $\|f_Y\|_\infty \le L$ and $\|f_Z\|_\infty \le L$, it is clear to see that

$$p_Y(\boldsymbol{\ell}) \le \|f_Y\|_\infty \kappa_{(1)}^{-d} \le L\kappa_{(1)}^{-d} \quad \text{and} \quad p_Z(\boldsymbol{\ell}) \le \|f_Z\|_\infty \kappa_{(1)}^{-d} \le L\kappa_{(1)}^{-d} \quad \text{for all } \boldsymbol{\ell}.$$

This gives

$$b_{(1)} = \max\{\|p_Y\|_2^2, \|p_Z\|_2^2\} \le L\kappa_{(1)}^{-d}. \tag{G.20}$$

Based on Lemma 3 of Arias-Castro et al. (2018), one can find a constant $C_1 > 0$ such that

$$\|p_Y - p_Z\|_2^2 \ge C_1 \kappa_{(1)}^{-d} \epsilon_{n_1,n_2}^2, \tag{G.21}$$

where ϵ_{n_1,n_2} is the lower bound for $\|f_Y - f_Z\|_{L_2}$. By combining (G.20) and (G.21), the condition of Proposition 8.1 is satisfied when

$$\kappa_{(1)}^{-d}\epsilon_{n_1,n_2}^2 \;\ge\; C_2 \frac{L^{1/2}\kappa_{(1)}^{-d/2}}{\alpha^{1/2}\beta}\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Equivalently,

$$\epsilon_{n_1,n_2} \;\ge\; C_3 \frac{L^{1/4}\kappa_{(1)}^{d/4}}{\alpha^{1/4}\beta^{1/2}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}.$$

Since $\kappa_{(1)} = \lfloor n_1^{2/(4s+d)} \rfloor$ and we assume $n_1 \le n_2$, the above inequality is further implied by

$$\epsilon_{n_1,n_2} \;\ge\; \frac{C_4}{\alpha^{1/4}\beta^{1/2}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{2s}{4s+d}},$$

where C_4 is a constant that may depend on s, d, L. This completes the proof of Proposition 8.3.

G.9 Proof of Theorem 8.2

The proof of Theorem 8.2 is similar to that of Theorem 8.1. First we verify that the permuted U-statistic U_n^π, which can be recalled from (8.32), has zero expectation. By the linearity of expectation, the problem

boils down to showing

$$\mathbb{E}_\pi[h_{\text{in}}\{(Y_1, Z_{\pi_1}), (Y_2, Z_{\pi_2}), (Y_3, Z_{\pi_3}), (Y_4, Z_{\pi_4})\}|\mathcal{X}_n] = 0.$$

Since $Y_1, \ldots, Y_4$ are constant under permutations, it further boils down to proving

$$\mathbb{E}_\pi[g_Z(Z_{\pi_1}, Z_{\pi_2}) + g_Z(Z_{\pi_3}, Z_{\pi_4}) - g_Z(Z_{\pi_1}, Z_{\pi_3}) - g_Z(Z_{\pi_2}, Z_{\pi_4})|\mathcal{X}_n] = 0.$$

In fact, this equality is clear by noting that $\mathbb{E}_\pi[g_Z(Z_{\pi_i}, Z_{\pi_j})]$ is invariant to the choice of $(i, j) \in \mathbf{i}_2^n$, which leads to $\mathbb{E}_\pi[U_n^\pi|\mathcal{X}_n] = 0$. Therefore we can focus on the simplified condition (8.4) to proceed.

The rest of the proof is split into two parts. In each part, we prove the following conditions separately,

$$\mathbb{E}_P[U_n] \geq 2\sqrt{\frac{2\text{Var}_P[U_n]}{\beta}} \quad \text{and} \tag{G.22}$$

$$\mathbb{E}_P[U_n] \geq 2\sqrt{\frac{2\mathbb{E}_P[\text{Var}_\pi\{U_n^\pi|\mathcal{X}_n\}]}{\alpha\beta}}. \tag{G.23}$$

We then complete the proof of Theorem 8.2 by noting that (G.22) and (G.23) imply the simplified condition (8.4).

Part 1. Verification of condition (G.22): This part verifies condition (G.22). The main ingredient of this part of the proof is the explicit variance formula of a U-statistic (e.g. page 12 of Lee, 1990). Following the notation of Lee (1990), we define $\check{\sigma}_i^2$ to be the variance of the conditional expectation by

$$\check{\sigma}_i^2 := \text{Var}_P[\mathbb{E}_P\{\overline{h}_{\text{in}}(x_1, \ldots, x_i, X_{i+1}, \ldots, X_4)\}] \quad \text{for } 1 \leq i \leq 4.$$

Then the variance of U_n is given by

$$\text{Var}_P[U_n] = \sum_{i=1}^{4} \binom{4}{i}\binom{n-4}{4-i}\binom{n}{4}^{-1} \check{\sigma}_i^2.$$

By the law of total variance, it can be seen that $\check{\sigma}_i^2 \leq \check{\sigma}_4^2$ for all $1 \leq i \leq 4$, which leads to an upper bound for $\text{Var}_P[U_n]$ as

$$\text{Var}_P[U_n] \leq C_1\frac{\check{\sigma}_1^2}{n} + C_2\frac{\check{\sigma}_4^2}{n^2}.$$

Now applying Jensen's inequality, repeatedly, yields

$$\check{\sigma}_4^2 \leq \mathbb{E}_P[\bar{h}_{\text{in}}^2(X_1, X_2, X_3, X_4)] \leq \mathbb{E}_P[h_{\text{in}}^2(X_1, X_2, X_3, X_4)] \leq C_3 \psi_2'(P).$$

Then by noting that $\check{\sigma}_1^2$ corresponds to the notation $\psi_1'(P)$, we have that

$$\text{Var}_P[U_{n_1,n_2}] \leq C_1 \frac{\psi_1'(P)}{n} + C_2 \frac{\psi_2'(P)}{n^2}.$$

Therefore condition (G.22) is satisfied by taking constant C sufficiently large in Theorem 8.2.

Part 2. Verification of condition (G.23): This part verifies condition (G.23). As mentioned in the main text, the permuted U-statistic U_n^π mimics the behavior of U_n under the null hypothesis. Hence one can expect that the variance of U_n^π is similarly bounded by $\psi_2'(P)n^{-2}$ up to some constant as $\psi_1'(P)$ becomes zero under the null hypothesis. To prove this statement, we first introduce some notation. Let us define a set of indices $\mathsf{J}_{\text{total}} := \{(i_1, i_2, i_3, i_4, i_1', i_2', i_3', i_4') \in \mathbb{N}_+^8 : (i_1, i_2, i_3, i_4) \in \mathbf{i}_4^n, (i_1', i_2', i_3', i_4') \in \mathbf{i}_4^n\}$ and let $\mathsf{J}_{\text{A}} := \{(i_1, i_2, i_3, i_4, i_1', i_2', i_3', i_4') \in \mathsf{J}_{\text{total}} : \#|\{i_1, i_2, i_3, i_4\} \cap \{i_1', i_2', i_3', i_4'\}| \leq 1\}$ and $\mathsf{J}_{\text{A}^c} := \{(i_1, i_2, i_3, i_4, i_1', i_2', i_3', i_4') \in \mathsf{J}_{\text{total}} : \#|\{i_1, i_2, i_3, i_4\} \cap \{i_1', i_2', i_3', i_4'\}| > 1\}$. By construction, it is clear that $\mathsf{J}_{\text{total}} = \mathsf{J}_{\text{A}} \cup \mathsf{J}_{\text{A}^c}$. To shorten the notation, we simply write

$$h_{\text{in}}(x_1, x_2, x_3, x_4) = h_{\text{in},Y}(y_1, y_2, y_3, y_4) h_{\text{in},Z}(z_1, z_2, z_3, z_4),$$

where $h_{\text{in},Y}(y_1, y_2, y_3, y_4) := g_Y(y_1, y_2) + g_Y(y_3, y_4) - g_Y(y_1, y_3) - g_Y(y_2, y_4)$ and $h_{\text{in},Z}(z_1, z_2, z_3, z_4) := g_Z(z_1, z_2) + g_Z(z_3, z_4) - g_Z(z_1, z_3) - g_Z(z_2, z_4)$. Since U_n^π is centered, our interest is in bounding $\mathbb{E}_P[\mathbb{E}_\pi\{(U_n^\pi)^2 | \mathcal{X}_n\}]$. Focusing on the conditional expectation inside, observe that

$$\begin{aligned}
\mathbb{E}_\pi[(U_n^\pi)^2 | \mathcal{X}_n] &= \frac{1}{n_{(4)}^2} \sum_{(i_1,\ldots,i_4') \in \mathsf{J}_{\text{total}}} h_{\text{in},Y}(Y_{i_1}, Y_{i_2}, Y_{i_3}, Y_{i_4}) h_{\text{in},Y}(Y_{i_1'}, Y_{i_2'}, Y_{i_3'}, Y_{i_4'}) \\
&\quad \times \mathbb{E}_\pi\left[h_{\text{in},Z}(Z_{\pi_{i_1}}, Z_{\pi_{i_2}}, Z_{\pi_{i_3}}, Z_{\pi_{i_4}}) h_{\text{in},Z}(Z_{\pi_{i_1'}}, Z_{\pi_{i_2'}}, Z_{\pi_{i_3'}}, Z_{\pi_{i_4'}}) | \mathcal{X}_n\right] \\
&= (I') + (II'),
\end{aligned}$$

where

$$\begin{aligned}
(I') &:= \frac{1}{n_{(4)}^2} \sum_{(i_1,\ldots,i_4') \in \mathsf{J}_{\text{A}}} h_{\text{in},Y}(Y_{i_1}, Y_{i_2}, Y_{i_3}, Y_{i_4}) h_{\text{in},Y}(Y_{i_1'}, Y_{i_2'}, Y_{i_3'}, Y_{i_4'}) \\
&\quad \times \mathbb{E}_\pi\left[h_{\text{in},Z}(Z_{\pi_{i_1}}, Z_{\pi_{i_2}}, Z_{\pi_{i_3}}, Z_{\pi_{i_4}}) h_{\text{in},Z}(Z_{\pi_{i_1'}}, Z_{\pi_{i_2'}}, Z_{\pi_{i_3'}}, Z_{\pi_{i_4'}}) | \mathcal{X}_n\right],
\end{aligned}$$

$$(II') := \frac{1}{n^2_{(4)}} \sum_{(i_1,\ldots,i'_4)\in \mathsf{J}_{\mathrm{A}^c}} h_{\mathrm{in},Y}(Y_{i_1},Y_{i_2},Y_{i_3},Y_{i_4}) h_{\mathrm{in},Y}(Y_{i'_1},Y_{i'_2},Y_{i'_3},Y_{i'_4}) \\ \times \mathbb{E}_\pi\big[h_{\mathrm{in},Z}(Z_{\pi_{i_1}},Z_{\pi_{i_2}},Z_{\pi_{i_3}},Z_{\pi_{i_4}}) h_{\mathrm{in},Z}(Z_{\pi_{i'_1}},Z_{\pi_{i'_2}},Z_{\pi_{i'_3}},Z_{\pi_{i'_4}})|\mathcal{X}_n\big].$$

We now claim that the first term $(I') = 0$, which is critical to obtain a faster rate n^{-2} rather than n^{-1} in the bound (8.21). However we have already proved in the second part of the proof of Theorem 8.1 that

$$\mathbb{E}_\pi\big[h_{\mathrm{in},Z}(Z_{\pi_{i_1}},Z_{\pi_{i_2}},Z_{\pi_{i_3}},Z_{\pi_{i_4}}) h_{\mathrm{in},Z}(Z_{\pi_{i'_1}},Z_{\pi_{i'_2}},Z_{\pi_{i'_3}},Z_{\pi_{i'_4}})|\mathcal{X}_n\big] = 0,$$

whenever $(i_1,\ldots,i'_4) \in \mathsf{J}_A$. This concludes $(I') = 0$ and so $\mathbb{E}_\pi[(U^\pi_n)^2|\mathcal{X}_n] = (II')$. To bound $\mathbb{E}_P[(II')]$, we make an observation that for any $1 \le i_1 \neq i_2, i'_1 \neq i'_2 \le n$,

$$\begin{aligned}
&\big|\mathbb{E}_P\big[g_Y(Y_{i_1},Y_{i_2}) g_Y(Y_{i'_1},Y_{i'_2}) \mathbb{E}_\pi\big\{g_Z(Z_{\pi_{i_1}},Z_{\pi_{i_2}}) g_Z(Z_{\pi_{i'_1}},Z_{\pi_{i'_2}})|\mathcal{X}_n\big\}\big]\big| \\
\overset{(i)}{=}\ &\big|\mathbb{E}_\pi\big[\mathbb{E}_P\big\{g_Y(Y_{i_1},Y_{i_2}) g_Y(Y_{i'_1},Y_{i'_2}) g_Z(Z_{\pi_{i_1}},Z_{\pi_{i_2}}) g_Z(Z_{\pi_{i'_1}},Z_{\pi_{i'_2}})|\pi\big\}\big]\big| \\
\overset{(ii)}{\le}\ &\frac{1}{2}\mathbb{E}_\pi\big[\mathbb{E}_P\big\{g^2_Y(Y_{i_1},Y_{i_2}) g^2_Z(Z_{\pi_{i_1}},Z_{\pi_{i_2}})|\pi\big\}\big] + \frac{1}{2}\mathbb{E}_\pi\big[\mathbb{E}_P\big\{g^2_Y(Y_{i'_1},Y_{i'_2}) g^2_Z(Z_{\pi_{i'_1}},Z_{\pi_{i'_2}})|\pi\big\}\big] \\
\overset{(iii)}{\le}\ &\psi'_2(P),
\end{aligned}$$

where (i) uses the law of total expectation, (ii) uses the basic inequality $xy \le x^2/2 + y^2/2$ and (iii) follows by the definition of $\psi'_2(P)$. Based on this observation, it is difficult to see that for any $(i_1,\ldots,i'_4) \in \mathsf{J}_{\mathrm{total}}$,

$$\big|\mathbb{E}_P[\mathbb{E}_\pi\{h_{\mathrm{in}}(X_{\pi_{i_1}},X_{\pi_{i_2}},X_{\pi_{i_3}},X_{\pi_{i_4}}) h_{\mathrm{in}}(X_{\pi_{i'_1}},X_{\pi_{i'_2}},X_{\pi_{i'_3}},X_{\pi_{i'_4}})|\mathcal{X}_n\}]\big| \le C_1\psi'_2(P).$$

Therefore, by counting the number of elements in $\mathsf{J}_{\mathrm{A}^c}$,

$$\begin{aligned}
\mathbb{E}_P[\mathrm{Var}_\pi\{U^\pi_n|\mathcal{X}_n\}] &= \mathbb{E}_P[(II')] \\
&\le C_2\psi'_2(P)\frac{1}{n^2_{(4)}} \sum_{(i_1,\ldots,i'_4)\in \mathsf{J}_{\mathrm{A}^c}} 1 \\
&\le C_3\frac{\psi'_2(P)}{n^2}.
\end{aligned}$$

Now by taking constant C in Theorem 8.2 sufficiently large, one may see that condition (G.23) is satisfied. This completes the proof of Theorem 8.2.

G.10 Proof of Proposition 8.4

To prove Proposition 8.4, it suffices to verify that the two inequalities (8.24) hold. Then the result follows by Theorem 8.2. To start with the first inequality in (8.24), we want to upper bound $\psi_1'(P)$ as $\psi_1'(P) \leq C_1\sqrt{b_{(2)}}\|p_{YZ} - p_Y p_Z\|_2^2$. A little algebra shows that

$$\begin{aligned}
&\mathbb{E}_P[\overline{h}_{\text{in}}(X_1, X_2, X_3, X_4)|X_2, X_3, X_4] - 4\|p_{YZ} - p_Y p_Z\|_2^2 \\
=\ & 2\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}\big[\mathbb{1}(Y_1 = k)\mathbb{1}(Z_1 = k') - p_{YZ}(k,k')\big]\big[p_{YZ}(k,k') - p_Y(k)p_Z(k')\big] \\
& -2\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}\big[\mathbb{1}(Y_1 = k) - p_Y(k)\big]p_Z(k')\big[p_{YZ}(k,k') - p_Y(k)p_Z(k')\big] \\
& -2\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}\big[\mathbb{1}(Z_1 = k') - p_Z(k')\big]p_Y(k)\big[p_{YZ}(k,k') - p_Y(k)p_Z(k')\big] \\
:=\ & 2(I) - 2(II) - 2(III) \quad \text{(say)}.
\end{aligned}$$

Then by recalling the definition of $\psi_1'(P)$ in (8.19) and based on the elementary inequality $(x_1+x_2+x_3)^2 \leq 3x_1^2 + 3x_2^2 + 3x_3^2$, we have

$$\psi_1'(P) \leq 12\mathbb{E}_P[(I)^2] + 12\mathbb{E}_P[(II)^2] + 12\mathbb{E}_P[(III)^2].$$

For convenience, we write $\Delta_{k,k'} := p_{YZ}(k,k') - p_Y(k)p_Z(k')$. Focusing on the first expectation in the above upper bound, the basic inequality $(x+y)^2 \leq x^2/ + y^2/2$ gives

$$\begin{aligned}
\mathbb{E}_P[(I)^2] \ &\leq\ \frac{1}{2}\mathbb{E}_P\bigg[\bigg\{\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}\mathbb{1}(Y_1 = k)\mathbb{1}(Z_1 = k')\Delta_{k,k'}\bigg\}^2\bigg] + \frac{1}{2}\bigg\{\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}p_{YZ}(k,k')\Delta_{k,k'}\bigg\}^2 \\
&\overset{(i)}{\leq}\ \frac{1}{2}\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}p_{YZ}(k,k')\Delta_{k,k'}^2 + \frac{1}{2}\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}p_{YZ}^2(k,k')\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}\Delta_{k,k'}^2 \\
&\overset{(ii)}{\leq}\ \frac{1}{2}\sqrt{\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}p_{YZ}^2(k,k')}\sqrt{\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}\Delta_{k,k'}^4} + \frac{1}{2}\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}p_{YZ}^2(k,k')\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}\Delta_{k,k'}^2 \\
&\overset{(iii)}{\leq}\ \sqrt{b_{(2)}}\|p_{YZ} - p_Y p_Z\|_2^2,
\end{aligned}$$

where (i) and (ii) use Cauchy-Schwarz inequality and the monotonicity of ℓ_p norm (specifically, $\ell_4 \leq \ell_2$). (iii) follows by the definition of $b_{(2)}$ in (8.23) and the fact that $\|p_{YZ}\|_2^2 \leq \|p_{YZ}\|_2$. Turning to the second

term (II), one may see that

$$\begin{aligned}\mathbb{E}_P[(II)^2] &\leq \frac{1}{2}\mathbb{E}_P\left[\left\{\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}\mathbb{1}(Y_1=k)p_Z(k')\Delta_{k,k'}\right\}^2\right] + \frac{1}{2}\left\{\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}p_Y(k)p_Z(k')\Delta_{k,k'}\right\}^2 \\ &= \frac{1}{2}(II)_a + \frac{1}{2}(II)_b \quad \text{(say)}.\end{aligned}$$

Using the fact that $\mathbb{1}(Y_1=k_1)\mathbb{1}(Y_1=k_2) = \mathbb{1}(Y_1=k_1)\mathbb{1}(k_1=k_2)$, we may upper bound $(II)_a$ by

$$\begin{aligned}\mathbb{E}_P[(II)_a] &= \sum_{k=1}^{d_1}p_Y(k)\left[\sum_{k'=1}^{d_2}p_Z(k')\Delta_{k,k'}\right]^2 \\ &\overset{(i)}{\leq} \sqrt{\sum_{k=1}^{d_1}p_Y^2(k)}\sqrt{\sum_{k=1}^{d_1}\left(\sum_{k'=1}^{d_2}p_Z(k')\Delta_{k,k'}\right)^4} \\ &\overset{(ii)}{\leq} \sqrt{\sum_{k=1}^{d_1}p_Y^2(k)}\sqrt{\sum_{k=1}^{d_1}\left(\sum_{k'=1}^{d_2}p_Z^2(k')\sum_{k''=1}^{d_2}\Delta_{k,k''}^2\right)^2} \\ &\overset{(iii)}{\leq} \sqrt{\sum_{k=1}^{d_1}p_Y^2(k)\sum_{k'=1}^{d_2}p_Z^2(k')}\sqrt{\sum_{k=1}^{d_1}\left(\sum_{k'=1}^{d_2}\Delta_{k,k'}^2\right)^2} \\ &\overset{(iv)}{\leq} \sqrt{b_{(2)}}\|p_{YZ}-p_Yp_Z\|_2^2,\end{aligned}$$

where both (i) and (ii) use Cauchy-Schwarz inequality, (iii) uses $\|p_Z\|_2^2 \leq \|p_Z\|_2$ and (iii) follows by the monotonicity of ℓ_p norm (specifically, $\ell_2 \leq \ell_1$) and the definition of $b_{(2)}$ in (8.23). The second term $(II)_b$ is bounded similarly by Cauchy-Schwarz inequality and $\|p_Y\|_2^2 \leq \|p_Y\|_2$ and $\|p_Z\|_2^2 \leq \|p_Z\|_2$. In particular,

$$\mathbb{E}_P[(II)_b] \leq \sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}p_Y^2(k)p_Z^2(k')\|p_{YZ}-p_Yp_Z\|_2^2 \leq \sqrt{b_{(2)}}\|p_{YZ}-p_Yp_Z\|_2^2.$$

By symmetric, $\mathbb{E}_P[(III)^2]$ is also upper bounded by $\sqrt{b_{(2)}}\|p_{YZ}-p_Yp_Z\|_2^2$. Hence, putting things together, we have $\psi_1'(P) \leq C_1\sqrt{b_{(2)}}\|p_{YZ}-p_Yp_Z\|_2^2$.

Next we show that the second inequality of (8.24), which is $\psi_{(2)}'(P) \leq C_2 b_{(2)}$, holds. By recalling the definition of $\psi_{(2)}'(P)$ in (8.19) and noting that $g_Y^2(Y_1,Y_2) = g_Y(Y_1,Y_2)$ and $g_Z^2(Z_1,Z_2) = g_Z(Z_1,Z_2)$, we shall see that

$$\mathbb{E}_P[g_Y(Y_1,Y_2)g_Z(Z_1,Z_2)] = \sum_{k=1}^{d_1}\sum_{k'=1}^{d_2}p_{YZ}^2(k,k') \leq b_{(2)},$$

$$\begin{aligned}
\mathbb{E}_P[g_Y(Y_1,Y_2)g_Z(Z_1,Z_3)] &= \sum_{k=1}^{d_1}\sum_{k'=1}^{d_2} p_{YZ}(k,k')p_Y(k)p_Z(k') \\
&\leq \frac{1}{2}\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2} p_{YZ}^2(k,k') + \frac{1}{2}\sum_{k=1}^{d_1}\sum_{k'=1}^{d_2} p_Y^2(k)p_Z^2(k') \leq b_{(2)}, \\
\mathbb{E}_P[g_Y(Y_1,Y_2)g_Z(Z_3,Z_4)] &= \sum_{k=1}^{d_1}\sum_{k'=1}^{d_2} p_Y^2(k)p_Z^2(k') \leq b_{(2)}.
\end{aligned}$$

Hence both conditions in (8.24) are satisfied under the assumption in Proposition 8.4. This concludes Proposition 8.4.

G.11 Proof of Proposition 8.5

As in the proof of Proposition 8.2, we properly construct a mixture distribution Q and a null distribution P_0 and apply Lemma G.0.2 to prove the result. To start we consider P_0 to be the product of the uniform discrete distributions given by

$$P_0(k_1,k_2) := p_Y(k_1)p_Z(k_2) = \frac{1}{d_1 d_2} \quad \text{for all } k_1 = 1,\ldots,d_1 \text{ and } k_2 = 1,\ldots,d_2.$$

Let $\widetilde{\zeta} = \{\widetilde{\zeta}_1,\ldots,\widetilde{\zeta}_{d_1}\}$ and $\widetilde{\xi} = \{\widetilde{\xi}_1,\ldots,\widetilde{\xi}_{d_2}\}$ be dependent Rademacher random variables uniformly distributed over $\mathcal{M}_{d_1}$ and $\mathcal{M}_{d_2}$, respectively, where $\mathcal{M}_{d_1}$ and $\mathcal{M}_{d_2}$ are hypercubes defined in (G.17). Assume that $\widetilde{\zeta}$ and $\widetilde{\xi}$ are independent. Let us denote the cardinality of $\mathcal{M}_{d_1}$ and $\mathcal{M}_{d_2}$ by N_1 and N_2, respectively. Given $\widetilde{\zeta} \in \mathcal{M}_{d_1}$ and $\widetilde{\xi} \in \mathcal{M}_{d_2}$, we define a distribution $p_{\widetilde{\zeta},\widetilde{\xi}}$ such that

$$p_{\widetilde{\zeta},\widetilde{\xi}}(k_1,k_2) := \frac{1}{d_1 d_2} + \delta \sum_{i_1=1}^{d_1}\sum_{i_2=1}^{d_2} \widetilde{\zeta}_{i_1}\widetilde{\xi}_{i_2}\mathbb{1}(k_1 = i_1)\mathbb{1}(k_2 = i_2),$$

where $\delta \leq 1/(d_1 d_2)$ and thus $\|p_{\widetilde{\zeta},\widetilde{\xi}}\|_2^2 \leq 2/(d_1 d_2)$. Since $\widetilde{\zeta} \in \mathcal{M}_{d_1}$ and $\widetilde{\xi} \in \mathcal{M}_{d_2}$, it is straightforward to check that

$$\begin{aligned}
&\sum_{k_1=1}^{d_1} p_{\widetilde{\zeta},\widetilde{\xi}}(k_1,k_2) = \frac{1}{d_2} + \delta\left\{\sum_{i_1=1}^{d_1}\widetilde{\zeta}_{i_1}\right\} \times \left\{\sum_{i_2=1}^{d_2}\widetilde{\xi}_{i_2}\mathbb{1}(k_2 = i_2)\right\} = \frac{1}{d_2}, \\
&\sum_{k_2=1}^{d_2} p_{\widetilde{\zeta},\widetilde{\xi}}(k_1,k_2) = \frac{1}{d_1} + \delta\left\{\sum_{i_1=1}^{d_1}\widetilde{\zeta}_{i_1}\mathbb{1}(k_1 = i_1)\right\} \times \left\{\sum_{i_2=1}^{d_2}\widetilde{\xi}_{i_2}\right\} = \frac{1}{d_1} \quad \text{and} \\
&\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2} p_{\widetilde{\zeta},\widetilde{\xi}}(k_1,k_2) = 1.
\end{aligned}$$

Therefore $p_{\widetilde{\zeta},\widetilde{\xi}}$ is a joint discrete distribution whose marginals are equivalent to those of the product distribution. Let us denote such distributions by $p_{\widetilde{\zeta}(1),\widetilde{\xi}(1)}, \ldots, p_{\widetilde{\zeta}(N_1),\widetilde{\xi}(N_2)}$. We then consider the uniform mixture Q given by

$$Q := \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} p_{\widetilde{\zeta}(i),\widetilde{\xi}(j)}.$$

Note that both $\{\widetilde{\zeta}_1, \ldots, \widetilde{\zeta}_{d_1}\}$ and $\{\widetilde{\xi}_1, \ldots, \widetilde{\xi}_{d_2}\}$ are negatively associated and these two sets are mutually independent by construction. Hence, following Proposition 7 of Dubhashi and Ranjan (1998), the pooled random variables $\{\widetilde{\zeta}_1, \ldots, \widetilde{\zeta}_{d_1}, \widetilde{\xi}_1, \ldots, \widetilde{\xi}_{d_2}\}$ are also negatively associated. Having this observation at hand, the remaining steps are exactly the same as those in the proof of Proposition 8.2. This together with Proposition 8.4 completes the proof of Proposition 8.5.

G.12 Proof of Proposition 8.6

The proof of Proposition 8.6 is based on Proposition 8.4 and similar to that of Proposition 8.3. By recalling the notation from Appendix G.8 and $\kappa_{(2)} = \lfloor n^{2/(4s+d_1+d_2)} \rfloor$, we define $\boldsymbol{H}_{\boldsymbol{\ell}_Y} := [(\boldsymbol{\ell}_Y - 1)/\kappa_{(2)}, \boldsymbol{\ell}_Y/\kappa_{(2)}]$ and $\boldsymbol{H}_{\boldsymbol{\ell}_Z} := [(\boldsymbol{\ell}_Z - 1)/\kappa_{(2)}, \boldsymbol{\ell}_Z/\kappa_{(2)}]$ where $\boldsymbol{\ell}_Y \in \{1, 2, \ldots, \kappa_{(2)}\}^{d_1}$ and $\boldsymbol{\ell}_Z \in \{1, 2, \ldots, \kappa_{(2)}\}^{d_2}$. Then we denote the joint and product discretized distributions by

$$p_{YZ}(\boldsymbol{\ell_Y}, \boldsymbol{\ell_Z}) := \int_{\boldsymbol{H}_{\boldsymbol{\ell}_Y} \times \boldsymbol{H}_{\boldsymbol{\ell}_Z}} f_{YZ}(t_Y, t_Z) dt_Y dt_Z \quad \text{and}$$
$$p_Y p_Z(\boldsymbol{\ell_Y}, \boldsymbol{\ell_Z}) := \int_{\boldsymbol{H}_{\boldsymbol{\ell}_Y} \times \boldsymbol{H}_{\boldsymbol{\ell}_Z}} f_Y(t_Y) f_Z(t_Z) dt_Y dt_Z.$$

Since both f_{YZ} and $f_Y f_Z$ are in Hölder's density class $\mathcal{P}^{(d_1+d_2,s)}_{\text{Hölder}}$ where $\|\!| f_Y f_Z |\!\|_\infty \le L$ and $\|\!| f_{YZ} |\!\|_\infty \le L$, it is clear to see that

$$p_{YZ}(\boldsymbol{\ell_Y}, \boldsymbol{\ell_Z}) \le \|\!| f_{YZ} |\!\|_\infty \kappa_{(2)}^{-(d_1+d_2)} \le L \kappa_{(2)}^{-(d_1+d_2)} \quad \text{and}$$
$$p_Y p_Z(\boldsymbol{\ell_Y}, \boldsymbol{\ell_Z}) \le \|\!| f_Y f_Z |\!\|_\infty \kappa_{(2)}^{-(d_1+d_2)} \le L \kappa_{(2)}^{-(d_1+d_2)} \quad \text{for all } \boldsymbol{\ell}_Y, \boldsymbol{\ell}_Z.$$

This leads to

$$b_{(2)} = \max\{\|p_{YZ}\|_2^2, \|p_Y p_Z\|_2^2\} \le L \kappa_{(2)}^{-(d_1+d_2)}. \tag{G.24}$$

Furthermore, based on Lemma 3 of Arias-Castro et al. (2018), one can find a constant $C_1 > 0$ such that

$$\|p_{YZ} - p_Y p_Z\|_2^2 \geq C_1 \kappa_{(2)}^{-(d_1+d_2)} \epsilon_n^2, \tag{G.25}$$

where ϵ_n is the lower bound for $\|f_{YZ} - f_Y f_Z\|_{L_2}$. By combining (G.24) and (G.25), the condition of Proposition 8.4 is satisfied when

$$\kappa_{(2)}^{-(d_1+d_2)} \epsilon_n^2 \;\geq\; C_2 \frac{L^{1/2} \kappa_{(2)}^{-(d_1+d_2)/2}}{\alpha^{1/2} \beta n}.$$

By putting $\kappa_{(2)} = \lfloor n^{2/(4s+d_1+d_2)} \rfloor$ and rearranging the terms, the above inequality is equivalent to

$$\epsilon_n \;\geq\; \frac{C_3}{\alpha^{1/4} \beta^{1/2}} \left(\frac{1}{n}\right)^{\frac{2s}{4s+d_1+d_2}},$$

where C_3 is a constant that may depend on s, d_1, d_2, L. This completes the proof of Proposition 8.6.

G.13 Proof of Proposition 8.7

The proof of Proposition 8.7 is standard based on Ingster's method in Lemma G.0.2. In particular we closely follow the proof of Theorem 1 in Arias-Castro et al. (2018) which builds on Ingster (1987). Let us start with the construction of a mixture distribution Q and a null distribution P_0.

• **Construction of Q and P_0.** Let f_Y and f_Z be the uniform density functions on $[0,1]^{d_1}$ and $[0,1]^{d_2}$, respectively. Then the density function of the baseline product distribution P_0 is defined by

$$f_0(y,z) := f_Y(y) f_Z(z) = 1 \quad \text{for all } (y,z) \in [0,1]^{d_1+d_2}.$$

We let $\varphi_Y : \mathbb{R}^{d_1} \mapsto \mathbb{R}$ and $\varphi_Z : \mathbb{R}^{d_2} \mapsto \mathbb{R}$ be infinitely differentiable functions supported on $[0,1]^{d_1}$ and $[0,1]^{d_2}$ respectively. Furthermore these two functions satisfy

$$\int_{[0,1]^{d_1}} \varphi_Y(y) dy = \int_{[0,1]^{d_2}} \varphi_Z(z) dz = 0 \quad \text{and}$$
$$\int_{[0,1]^{d_1}} \varphi_Y^2(y) dy = \int_{[0,1]^{d_2}} \varphi_Z^2(z) dz = 1.$$

For $\boldsymbol{i} \in \mathbb{Z}^{d_1}$, $\boldsymbol{j} \in \mathbb{Z}^{d_2}$ and a positive integer κ, we write $\varphi_{Y,\boldsymbol{i}}(x) = \kappa^{d_1/2} \varphi_Y(\kappa x - \boldsymbol{i} + 1)$ and $\varphi_{Z,\boldsymbol{j}}(x) = \kappa^{d_2/2} \varphi_Z(\kappa x - \boldsymbol{j} + 1)$ where $\varphi_{Y,\boldsymbol{i}}$ and $\varphi_{Z,\boldsymbol{j}}$ are supported on $[(\boldsymbol{i}-1)/\kappa, \boldsymbol{i}/\kappa]$ and $[(\boldsymbol{j}-1)/\kappa, \boldsymbol{j}/\kappa]$. By construction,

it can be seen that

$$\int_{[0,1]^{d_1}} \varphi_{Y,\boldsymbol{i}}^2(y)dy = \int_{[0,1]^{d_2}} \varphi_{Z,\boldsymbol{j}}^2(z)dz = 1,$$
$$\int_{[0,1]^{d_1}} \varphi_{Y,\boldsymbol{i}}(y)dy = \int_{[0,1]^{d_2}} \varphi_{Z,\boldsymbol{j}}(z)dz = 0 \quad \text{and}$$
$$\int_{[0,1]^{d_1}} \varphi_{Y,\boldsymbol{i}}(y)\varphi_{Y,\boldsymbol{i}'}(y)dy = \int_{[0,1]^{d_2}} \varphi_{Z,\boldsymbol{j}}(z)\varphi_{Z,\boldsymbol{j}'}(z)dz = 0,$$

for $\boldsymbol{i} \neq \boldsymbol{i}'$ and $\boldsymbol{j} \neq \boldsymbol{j}'$. We denote by $\zeta_{\boldsymbol{k}} \in \{0,1\}$ an i.i.d. sequence of Rademacher variables where $\boldsymbol{k} := (\boldsymbol{i},\boldsymbol{j}) \in [\kappa]^{d_1+d_2}$. Now for $\rho > 0$ specified later, let us define the density function of a mixture distribution Q by

$$f_\zeta(y,z) := f_0(y,z) + \rho \sum_{\boldsymbol{k}\in[\kappa]^{d_1+d_2}} \zeta_{\boldsymbol{k}} \varphi_{Y,\boldsymbol{i}}(y)\varphi_{Z,\boldsymbol{j}}(z).$$

By letting ρ such that $\rho\kappa^{(d_1+d_2)/2}\|\varphi_{Y,Z}\|_\infty \leq 1$ where $\varphi_{Y,Z}(y,z) := \varphi_Y(y)\varphi_Z(z)$, it is seen that f_ζ is a proper density function supported on $[0,1]^{d_1+d_2}$ such that

$$\int_{[0,1]^{d_1}} f_\zeta(y,z)dy = \int_{[0,1]^{d_2}} f_\zeta(y,z)dz = \int_{[0,1]^{d_1+d_2}} f_\zeta(y,z)dydz = 1.$$

Therefore f_ζ has the same marginal distributions as the product distribution f_0. Furthermore when $\rho\kappa^{(d_1+d_2)/2+s}M/L \leq 1$ where $M := \max\left\{4\|\varphi_{Y,Z}^{(\lfloor s\rfloor)}\|_\infty, 2\|\varphi_{Y,Z}^{(\lfloor s\rfloor+1)}\|_\infty\right\}$, it directly follows from the proof of Theorem 1 in Arias-Castro et al. (2018) that $f_\zeta \in \mathcal{P}_{\text{Hölder}}^{(d_1+d_2,s)}$. Having these two densities f_0 and f_ζ such that

$$\|f_\zeta - f_0\|_{L_2}^2 = \rho^2\kappa^{d_1+d_2} = \epsilon_n^2,$$

we next compute $\mathbb{E}_{P_0}[L_n^2]$.

• **Calculation of $\mathbb{E}_{P_0}[L_n^2]$.** By recalling that $f_0(y,z) = 1$ for $(y,z) \in [0,1]^{d_1+d_2}$, let us start by writing L_n^2 as

$$L_n^2 = \frac{1}{2^{2\kappa^{d_1+d_2}}} \sum_{\zeta,\zeta'\in\{-1,1\}^{\kappa^{d_1+d_2}}} \prod_{i=1}^n f_\zeta(Y_i,Z_i)f_{\zeta'}(Y_i,Z_i).$$

We then use the orthonormal property of $\varphi_{Y,\boldsymbol{i}}$ and $\varphi_{Z,\boldsymbol{j}}$ to see that

$$\begin{aligned}
\mathbb{E}_{P_0}[L_n^2] &= \frac{1}{2^{2\kappa^{d_1+d_2}}} \sum_{\zeta,\zeta' \in \{-1,1\}^{\kappa^{d_1+d_2}}} \prod_{i=1}^{n} \mathbb{E}_0 \left[1 + \rho^2 \sum_{\boldsymbol{k} \in [\kappa]^{d_1+d_2}} \zeta_{\boldsymbol{k}} \zeta'_{\boldsymbol{k}} \varphi^2_{Y,\boldsymbol{i}}(Y_i) \varphi^2_{Z,\boldsymbol{j}}(Z_i) \right] \\
&= \frac{1}{2^{2\kappa^{d_1+d_2}}} \sum_{\zeta,\zeta' \in \{-1,1\}^{\kappa^{d_1+d_2}}} \left[1 + \rho^2 \sum_{\boldsymbol{k} \in [\kappa]^{d_1+d_2}} \zeta_{\boldsymbol{k}} \zeta'_{\boldsymbol{k}} \right]^n \\
&\leq \mathbb{E}_{\zeta,\zeta'} \left[e^{n\rho^2 \langle \zeta, \zeta' \rangle} \right],
\end{aligned}$$

where the last inequality uses $(1+x)^n \leq e^{nx}$. Based on the independence among the components of ζ and ζ', we further observe that

$$\mathbb{E}_{\zeta,\zeta'} \left[e^{n\rho^2 \langle \zeta, \zeta' \rangle} \right] = \left\{ \cosh(n\rho^2) \right\}^{\kappa^{d_1+d_2}} \leq \exp\left(\kappa^{d_1+d_2} n^2 \rho^4 / 2 \right)$$

where the last inequality follows by $\cosh(x) \leq e^{x^2/2}$ for all $x \in \mathbb{R}$.

• **Completion of the proof.** We invoke Lemma G.0.2 to finish the proof. From the previous step, we know that

$$\mathbb{E}_{P_0}[L_n^2] \leq \exp\left(\kappa^{d_1+d_2} n^2 \rho^4 / 2 \right).$$

Therefore the condition in Lemma G.0.2 is fulfilled when

$$\kappa^{d_1+d_2} n^2 \rho^4 \leq 2 \log\{1 + 4(1-\alpha-\beta)^2\}.$$

Now by setting $\kappa = \lfloor n^{2/(4s+d_1+d_2)} \rfloor$ and $\rho = cn^{-(2s+d_1+d_2)/(4s+d_1+d_2)}$, the above condition is further implied by

$$c \leq 2 \log\{1 + 4(1-\alpha-\beta)^2\}.$$

Previously we also use the assumptions that $\rho \kappa^{(d_1+d_2)/2} \| \varphi_{Y,Z} \|_\infty \leq 1$ and $\rho \kappa^{(d_1+d_2)/2+s} M/L \leq 1$. These are satisfied by taking c sufficiently small. This means that when

$$\epsilon_n \leq c e^{-2s/(4s+d_1+d_2)},$$

for a small $c > 0$, the minimax type II error is less than β. Therefore, combined with Proposition 8.6, we complete the proof of Proposition 8.7.

G.14 Proof of Theorem 8.3

We continue the proof of Theorem 8.3 from the last line of (8.30). First we view $\widetilde{U}^{\pi,L,\zeta}_{n_1,n_2}$ as a quadratic form of ζ conditional on π and L. We then borrow the proof of Hanson–Wright inequality (see e.g. Rudelson and Vershynin, 2013; Vershynin, 2018) to proceed. To do so, let us denote $a_{k_1,k_2}(\pi, L) = h_{\mathrm{ts}}(X_{\pi_{k_1}}, X_{\pi_{k_2}}; X_{\pi_{n_1+\ell_{k_1}}}, X_{\pi_{n_1+\ell_{k_2}}})$ for $1 \le k_1 \ne k_2 \le n$ and $a_{k_1,k_2}(\pi, L) = 0$ for $1 \le k_1 = k_2 \le n$. Let $\mathbf{A}_{\pi,L}$ be the $n \times n$ matrix whose elements are $a_{k_1,k_2}(\pi, L)$. By following the proof of Theorem 1.1 in Rudelson and Vershynin (2013), we can obtain

$$e^{-\lambda t}\mathbb{E}_{\pi,L,\zeta}\big[\exp\big(\lambda\widetilde{U}^{\pi,L,\zeta}_{n_1,n_2}\big)|\mathcal{X}_n\big] \le \mathbb{E}_{\pi,L}\big[\exp\big(-\lambda t + C\lambda^2\|\mathbf{A}_{\pi,L}\|_F^2\big)\big],$$

which holds for $0 \le \lambda \le c/\|\mathbf{A}_{\pi,L}\|_{\mathrm{op}}$. Here, $\|\mathbf{A}_{\pi,L}\|_F$ and $\|\mathbf{A}_{\pi,L}\|_{\mathrm{op}}$ denote the Frobenius norm and the operator norm of $\mathbf{A}_{\pi,L}$, respectively. By optimizing over $0 \le \lambda \le c/\|A_{\pi,L}\|_{\mathrm{op}}$, we have that

$$\mathbb{P}_\pi(U^\pi_{n_1,n_2} \ge t \mid \mathcal{X}_n) \le \mathbb{E}_{\pi,L}\left[\exp\left\{-C_1 \min\left(\frac{t^2}{\|\mathbf{A}_{\pi,L}\|_F^2}, \frac{t}{\|\mathbf{A}_{\pi,L}\|_{\mathrm{op}}}\right)\right\}\right].$$

The proof of Theorem 8.3 is completed by noting that $\|\mathbf{A}_{\pi,L}\|_{\mathrm{op}} \le \|\mathbf{A}_{\pi,L}\|_F \le C_2\Sigma_{n_1,n_2}$.

G.15 Proof of Corollary 8.5.1

Note that the following equality holds:

$$\begin{aligned}\sum_{(i,j)\in\mathbf{i}_2^n} \widetilde{\zeta}_i\widetilde{\zeta}_j\,(a_{i,j} - \overline{a}) \;=\; & \frac{1}{4(n-1)(n-2)} \sum_{(i_1,i_2,i_3,i_4)\in\mathbf{i}_4^n} \Big\{ \\ & (\widetilde{\zeta}_{i_1}\widetilde{\zeta}_{i_3} + \widetilde{\zeta}_{i_2}\widetilde{\zeta}_{i_4} - \widetilde{\zeta}_{i_1}\widetilde{\zeta}_{i_4} - \widetilde{\zeta}_{i_2}\widetilde{\zeta}_{i_3})\,(a_{i_1,i_3} + a_{i_2,i_4} - a_{i_1,i_4} - a_{i_2,i_3}) \Big\},\end{aligned}$$

which can be verified by expanding the summation on the right-hand side. We also note that $\{\widetilde{\zeta}_1, \ldots, \widetilde{\zeta}_n\} \overset{d}{=} \{b_{\pi_1}, \ldots, b_{\pi_n}\}$ where $b_i = 1$ for $i = 1, \ldots, n/2$ and $b_i = -1$ for $i = n/2+1, \ldots, n$. Therefore, we can apply Theorem 8.4 with the bound of Σ_n^2 in (8.36). To be clear, $a_{i,j}$ does not need to be symmetric in its arguments. Theorem 8.4 still holds as long as g_Z is symmetric (g_Y is not necessarily symmetric), which is the case for this application. Alternatively, one can work with the symmetrized version of $a_{i,j}$, i.e. $\widetilde{a}_{i,j} := (a_{i,j} + a_{j,i})/2$ by observing that $\overline{a} = \overline{\widetilde{a}} := n_{(2)}^{-1}\sum_{(i_1,i_2)\in\mathbf{i}_2^n} \widetilde{a}_{i_1,i_2}$ and

$$\sum_{(i,j)\in\mathbf{i}_2^n} \widetilde{\zeta}_i\widetilde{\zeta}_j\,(a_{i,j} - \overline{a}) = \sum_{(i,j)\in\mathbf{i}_2^n} \widetilde{\zeta}_i\widetilde{\zeta}_j\left(\widetilde{a}_{i,j} - \overline{\widetilde{a}}\right).$$

This completes the proof of Corollary 8.5.1.

G.16 Proof of Theorem 8.5

Continuing our discussion from the main text, we prove Theorem 8.5 in two steps. In the first step, we replace two independent permutations π, π' in $\widetilde{U}_n^{\pi,\pi',\zeta}$ with their i.i.d. counterparts $\widetilde{\pi}, \widetilde{\pi}'$. Once this decoupling step is done, the resulting statistic can be viewed as a usual degenerate U-statistic of i.i.d. random variables conditional on $\mathcal{X}_n$. This means that we can apply the concentration inequalities for degenerate U-statistics in De la Pena and Giné (1999) to finish the proof. This shall be done in the second step. For notational convenience, we write

$$\begin{aligned} & h_{\pi,\pi'}(i_1, i_2, i_1+m, i_2+m) \\ := \; & h_{\text{in}}\{(Y_{\pi'_{i_1}}, Z_{\pi_{i_1}}), (Y_{\pi'_{i_2}}, Z_{\pi_{i_2}}), (Y_{\pi'_{i_2+m}}, Z_{\pi_{\pi_{i_2+m}}}), (Y_{\pi'_{i_1+m}}, Z_{\pi_{i_1+m}})\}, \end{aligned} \tag{G.26}$$

throughout this proof.

1. Decoupling. We start with the decoupling part. Let $\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}$ be defined similarly as $\widetilde{U}_n^{\pi,\pi',\zeta}$ but with decoupled permutations $(\widetilde{\pi}, \widetilde{\pi}')$ instead of the original permutations (π, π'). Our goal here is to bound

$$\mathbb{E}_{\pi,\pi',\zeta}\big[\Psi\big(\lambda \widetilde{U}_n^{\pi,\pi',\zeta}\big)|\mathcal{X}_n\big] \le \mathbb{E}_{\widetilde{\pi},\widetilde{\pi}',\zeta}\big[\Psi\big(C_n \lambda \widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}\big)|\mathcal{X}_n\big], \tag{G.27}$$

where $c < C_n < C$ is some deterministic sequence depending on n with some positive constants $c, C > 0$. The way how we associate the original statistic $\widetilde{U}_n^{\pi,\pi',\zeta}$ with the decoupled couterpart $\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}$ is as follows. First, we construct a random subset K of $\{1, \ldots, n\}$ such that $\{\pi\}_{i\in K}$ and $\{\widetilde{\pi}\}_{i\in K}$ have the same distribution so that two test statistics based on $\{\pi\}_{i\in K}$ and $\{\widetilde{\pi}\}_{i\in K}$, respectively, shall have the same distribution. The remainder of the proof is devoted to replacing the subset of permutations $\{\pi_i\}_{i\in K}$ and $\{\widetilde{\pi}\}_{i\in K}$ with the entire set of permutations $\{\pi_i\}_{i=1}^n$ and $\{\widetilde{\pi}_i\}_{i=1}^n$. As far as we know, this idea was first employed by Duembgen (1998) to decouple the simple linear permuted statistic.

Let us make this decoupling idea more precise. To do so, we define K to be a random subset of $\{1, \ldots, n\}$ independent of everything else except $\widetilde{\pi}$. Specifically, we assume that the conditional distribution of K given $\widetilde{\pi}$ has the uniform distribution on the set of all $J \in \{1, \ldots, n\}$ such that

$$\{\widetilde{\pi}_i : 1 \le i \le n\} = \{\widetilde{\pi}_i : i \in J\} \quad \text{and} \quad \#|\{\widetilde{\pi}_i : 1 \le i \le n\}| = \#|J|,$$

where $\#|A|$ denotes the cardinality of a set A. Then as noted in Duembgen (1998), $\{\pi_i\}_{i\in K} \overset{d}{=} \{\widetilde{\pi}_i\}_{i\in K}$ follows. In the same way, define another random subset K' of $\{1, \ldots, n\}$ only depending on $\widetilde{\pi}'$ such that $\{\pi'_i\}_{i\in K'} \overset{d}{=} \{\widetilde{\pi}'_i\}_{i\in K'}$; note that, by construction, K and K' are independent. Furthermore, we let $\mathcal{B}_{K,n}(i_1, i_2, i_1+m, i_2+m)$ be the event such that all of $\{i_1, i_2, i_1+m, i_2+m\}$ are in the random subset K.

Then, as $\{\pi_i\}_{i\in K} \stackrel{d}{=} \{\widetilde{\pi}_i\}_{i\in K}$ and $\{\pi'_i\}_{i\in K'} \stackrel{d}{=} \{\widetilde{\pi}'_i\}_{i\in K'}$, we may observe that

$$\widetilde{U}_n^{\pi,\pi',\zeta}(K,K') \;\stackrel{d}{=}\; \widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}(K,K'), \tag{G.28}$$

where

$$\begin{aligned}
\widetilde{U}_n^{\pi,\pi',\zeta}(K,K') :=\; & \frac{1}{m_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^m} \zeta_{i_1}\zeta_{i_2}\zeta_{i_1+m}\zeta_{i_2+m} h_{\pi,\pi'}(i_1,i_2,i_1+m,i_2+m)\times \\
& \mathbb{1}\{\mathcal{B}_{K,n}(i_1,i_2,i_1+m,i_2+m)\}\mathbb{1}\{\mathcal{B}_{K',n}(i_1,i_2,i_1+m,i_2+m)\}, \\
\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}(K,K') :=\; & \frac{1}{m_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^m} \zeta_{i_1}\zeta_{i_2}\zeta_{i_1+m}\zeta_{i_2+m} h_{\widetilde{\pi},\widetilde{\pi}'}(i_1,i_2,i_1+m,i_2+m)\times \\
& \mathbb{1}\{\mathcal{B}_{K,n}(i_1,i_2,i_1+m,i_2+m)\}\mathbb{1}\{\mathcal{B}_{K',n}(i_1,i_2,i_1+m,i_2+m)\}.
\end{aligned}$$

Next we calculate the probability of $\mathcal{B}_{K,n}(i_1,i_2,i_1+m,i_2+m)$. By symmetry, we may assume that $i_1 = 1, i_2 = 2, i_1+m = 3, i_2+m = 4$. In fact, this probability is the same as the probability that all of the first four urns are not empty when one throws n balls independently into n urns (here, each urn is equally likely to be selected). Based on the inclusion–exclusion formula, this probability can be computed as

$$B_n := \mathbb{P}\{\mathcal{B}_{K,n}(1,2,3,4)\} = 1 - 4\left(1-\frac{1}{n}\right)^n + 6\left(1-\frac{2}{n}\right)^n - 4\left(1-\frac{3}{n}\right)^n + \left(1-\frac{4}{n}\right)^n.$$

Indeed, B_n is monotone increasing for all $n \geq 4$. Hence we have that $\ell \leq B_n \leq u$ for any $n \geq 4$ where $\ell = 1 - 4(3/4)^4 + 6\,(1/2)^4 - 4\,(1/4)^4 = 0.09375$ and $u = 1 - 4e^{-1} + 6e^{-2} - 4e^{-3} + e^{-4} \approx 0.1597$. In the next step, we replace the subset of permutations $\{\pi_i\}_{i\in K}$ with the entire set of permutations $\{\pi_i\}_{i=1}^n$ as follows:

$$\begin{aligned}
\mathbb{E}_{\pi,\pi',\zeta}\big[\Psi\big(\lambda\widetilde{U}_n^{\pi,\pi',\zeta}\big)|\mathcal{X}_n\big] &\stackrel{(i)}{\leq} \mathbb{E}_{\pi,\pi',\zeta,K,K'}\big[\Psi\big\{B_n^{-2}\lambda\widetilde{U}_n^{\pi,\pi',\zeta}(K,K')\big\}|\mathcal{X}_n\big] \\
&\stackrel{(ii)}{=} \mathbb{E}_{\widetilde{\pi},\widetilde{\pi}',\zeta,K,K'}\big[\Psi\big\{B_n^{-2}\lambda\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}(K,K')\big\}|\mathcal{X}_n\big] \\
&\stackrel{(iii)}{\leq} \mathbb{E}_{\widetilde{\pi},\widetilde{\pi}',\zeta}\big[\Psi\big(B_n^{-2}\lambda\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}\big)|\mathcal{X}_n\big],
\end{aligned}$$

where (i) holds by Jensen's inequality with $\mathbb{E}_{K,K'}[\widetilde{U}_n^{\pi,\pi',\zeta}(K,K')] = B_n^2\widetilde{U}_n^{\pi,\pi',\zeta}$, (ii) is due to the relationship (G.28) and (iii) uses Jensen's inequality again with

$$\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}(K,K') = \mathbb{E}_\zeta\big[\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta} \;\big|\; \{\zeta_i\}_{i\in K}, \{\zeta_i\}_{i\in K'}, K, K', \mathcal{X}_n, \widetilde{\pi}, \widetilde{\pi}'\big].$$

This proves the decoupling inequality in (G.27).

2. Concentration. Having established the decoupled bound in (G.27), we are now ready to obtain the main result of Theorem 8.5. This part of the proof is largely based on Chapter 4.1.3 of De la Pena and Giné (1999). Recall that

$$\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta} \stackrel{d}{=} \frac{1}{m_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^m} \zeta_{i_1}\zeta_{i_2} h_{\widetilde{\pi},\widetilde{\pi}'}(i_1,i_2,i_1+m,i_2+m)$$

and $h_{\widetilde{\pi},\widetilde{\pi}'}(i_1,i_2,i_1+m,i_2+m)$ is given in (G.26). Let us write $\mathbf{Q}_{i_1} = ((Y_{\widetilde{\pi}'_{i_1}}, Z_{\widetilde{\pi}_{i_1}}), (Y_{\widetilde{\pi}'_{i_1+m}}, Z_{\widetilde{\pi}_{i_1+m}}))$ and $\mathbf{Q}_{i_2} = ((Y_{\widetilde{\pi}'_{i_2}}, Z_{\widetilde{\pi}_{i_2}}), (Y_{\widetilde{\pi}'_{i_2+m}}, Z_{\widetilde{\pi}_{i_2+m}}))$, which are random vectors with four main components. Note that $\mathbf{Q}_1,\ldots,\mathbf{Q}_m$ are independent and identically distributed conditional on $\mathcal{X}_n$. Define

$$h(\mathbf{Q}_{i_1},\mathbf{Q}_{i_2}) := h_{\widetilde{\pi},\widetilde{\pi}'}(i_1,i_2,i_1+m,i_2+m).$$

Then $\widetilde{U}_n^{\widetilde{\pi},\widetilde{\pi}',\zeta}$ can be viewed as a randomized U-statistic with the bivariate kernel $h(\mathbf{Q}_{i_1},\mathbf{Q}_{i_2})$. To summarize, we have established that

$$\mathbb{E}_\pi[\Psi(\lambda U_n^\pi)|\mathcal{X}_n] \;\le\; \mathbb{E}_{\zeta,Q}\left[\Psi\left(B_n^{-2}\lambda \frac{1}{m_{(2)}} \sum_{(i_1,i_2)\in \mathbf{i}_2^m} \zeta_{i_1}\zeta_{i_2} h(\mathbf{Q}_{i_1},\mathbf{Q}_{i_2})\right) \,\middle|\, \mathcal{X}_n\right].$$

Here, by letting $h^*(\mathbf{Q}_{i_1},\mathbf{Q}_{i_2}) = h(\mathbf{Q}_{i_1},\mathbf{Q}_{i_2})/2 + h(\mathbf{Q}_{i_2},\mathbf{Q}_{i_1})/2$, we may express the right-hand side of the above inequality with the symmetrized kernel as

$$\mathbb{E}_{\zeta,Q}\left[\Psi\left(B_n^{-2}\lambda \frac{2}{m_{(2)}} \sum_{1\le i_1<i_2\le m} \zeta_{i_1}\zeta_{i_2} h^*(\mathbf{Q}_{i_1},\mathbf{Q}_{i_2})\right) \,\middle|\, \mathcal{X}_n\right].$$

The rest of the proof follows exactly the same line of that of Theorem 4.1.12 in De la Pena and Giné (1999) based on (i) Chernoff bound, (ii) convex modification, (iii) Bernstein's inequality, (iv) hypercontractivity of Rademacher chaos variables and (v) Hoeffding's average (Hoeffding, 1963). In the end, we obtain

$$\mathbb{P}_\pi(nU_n^\pi \ge t \,|\mathcal{X}_n) \;\le\; C_1 \exp\left(-\lambda t^{2/3} + C_2\lambda^3\Lambda_n^2 + \frac{16C_2^2\Lambda_n^2 M_n^2\lambda^6}{n-(16/3)C_2M_n^2\lambda^3}\right),$$

for $n > (4/3)C_2M_n^2\lambda^3$, which corresponds to Equation (4.1.27) of De la Pena and Giné (1999). We complete the proof of Theorem 8.5 by optimizing the right-hand side over λ as detailed in De la Pena and Giné (1999).

G.17 Proof of Proposition 8.8

The proof of this result is motivated by Ingster (2000); Arias-Castro et al. (2018) and follows similarly as theirs. First note that type I error control of the adaptive test is trivial by the union bound. Hence we focus

on the type II error control. Note that by construction

$$\left(\frac{n_1}{\log\log n_1}\right)^{\frac{2}{4s+d}} \leq 2^{\gamma_{\max}}.$$

Therefore there exists an integer $j \in \{1, \ldots, \gamma_{\max}\}$ such that

$$2^{j-1} < \left(\frac{n_1}{\log\log n_1}\right)^{\frac{2}{4s+d}} \leq 2^j.$$

We take such j and define $\kappa^* := 2^j \in \mathsf{K}$. In the rest of the proof, we show that under the given condition, $\phi_{\kappa^*,\alpha/\gamma_{\max}}$ has the type II error at most β. If this is the case, then the proof is completed since $\mathbb{P}_P(\phi_{\text{adapt}} = 0) \leq \mathbb{P}_P(\phi_{\kappa^*,\alpha/\gamma_{\max}} = 0) \leq \beta$. To this end, let us start by improving Proposition 8.1 based on Lemma G.0.1. Using (8.14) and Lemma G.0.1, one can verify that Proposition 8.1 holds if

$$\|p_Y - p_Z\|_2^2 \geq \frac{C}{\beta}\log\left(\frac{1}{\alpha}\right)\frac{\sqrt{b_{(1)}}}{n_1}, \tag{G.29}$$

for some large constant $C > 0$ and $n_1 \asymp n_2$. Hence the multinomial test $\phi_{\kappa^*,\alpha/\gamma_{\max}}$ has the type II error at most β if condition (G.29) is fulfilled by replacing α with $\alpha/\gamma_{\max}$. Following the proof of Proposition 8.3 but with κ^* instead of $\kappa_{(1)}$, we can see that

$$b_{(1)} = \max\{\|p_Y\|_2^2, \|p_Z\|_2^2\} \leq L(\kappa^*)^{-d} \quad \text{and}$$
$$\|p_Y - p_Z\|_2^2 \geq C_1(s,d,L)(\kappa^*)^{-d}\epsilon_{n_1,n_2}^2.$$

Therefore condition (G.29) with $\alpha/\gamma_{\max}$ is satisfied when

$$\epsilon_{n_1,n_2}^2 \geq \frac{C_2(s,d,L)}{\beta}\log\left(\frac{\gamma_{\max}}{\alpha}\right)\frac{L^{1/2}\kappa^{*d/2}}{n_1}.$$

Based on the definition of $\gamma_{\max}$ and κ^*, the above inequality is further implied by

$$\epsilon_{n_1,n_2}^2 \geq C(s,d,L,\alpha,\beta)\left(\frac{\log\log n_1}{n_1}\right)^{\frac{4s}{4s+d}}.$$

This completes the proof of Proposition 8.8.

G.18 Proof of Proposition 8.9

The proof is almost identical to that of Proposition 8.8 once we establish the following lemma which is an improvement of Proposition 8.4.

Lemma G.0.3 (Multinomial independence testing)**.** *Let Y and Z be multinomial random vectors in $\mathbb{S}_{d_1'}$ and $\mathbb{S}_{d_2'}$, respectively. Consider the multinomial problem setting in Proposition 8.4 with an additional assumption that $n \geq C_1 d_1' d_2'$ for some positive constant $C_1 > 0$. Suppose that under the alternative hypothesis,*

$$\|p_{YZ} - p_Y p_Z\|_2 \geq \frac{C_2}{\beta^{1/2}} \sqrt{\log\left(\frac{1}{\alpha}\right)} \frac{b_{(2)}^{1/4}}{n^{1/2}},$$

for a sufficiently large $C_2 > 0$. Then the permutation test in Proposition 8.4 has the type II error at most β.

Proof. Following the proofs of Lemma G.0.1 and Proposition 8.4, we only need to show that the $1 - \beta/2$ quantile of the permutation critical value $c_{1-\alpha,n}$ of U_n, denoted by $q_{1-\beta/2,n}$, is bounded as

$$q_{1-\beta/2,n} \leq \frac{C_3}{\beta} \log\left(\frac{1}{\alpha}\right) \frac{b_{(2)}^{1/2}}{n}. \tag{G.30}$$

To establish this result, we first use the concentration bound in Theorem 8.5 to have

$$c_{1-\alpha,n} \leq C_4 \max\left\{ \frac{\Lambda_n}{n} \log\left(\frac{1}{\alpha}\right), \frac{1}{n^{3/2}} \log\left(\frac{1}{\alpha}\right) \right\},$$

where we use the fact that $M_n \leq 1$ and $\alpha \leq 1/2$. Hence, by Markov's inequality as in Lemma G.0.1, it can be seen that the quantile $q_{1-\beta/2,n}$ is bounded by

$$q_{1-\beta/2,n} \leq C_4 \max\left\{ \frac{\sqrt{2\mathbb{E}[\Lambda_n^2]}}{\beta^{1/2} n} \log\left(\frac{1}{\alpha}\right), \frac{1}{n^{3/2}} \log\left(\frac{1}{\alpha}\right) \right\}.$$

On the other hand, one can easily verify that

$$\mathbb{E}_P[\Lambda_n^2] = \frac{1}{n^4} \sum_{1 \leq i_1, i_2 \leq n} \sum_{1 \leq j_1, j_2 \leq n} \mathbb{E}\big[g_Y^2(Y_{i_1}, Y_{i_2}) g_Z^2(Z_{j_1}, Z_{j_2})\big] \leq b_{(2)} + \frac{C_5}{n}.$$

Furthermore, Cauchy-Schwarz inequality shows that

$$b_{(2)} = \max\{\|p_{YZ}\|_2^2, \|p_Y p_Z\|_2^2\} \geq \frac{1}{d_1' d_2'} \geq \frac{C_1}{n}, \tag{G.31}$$

where the last inequality uses the assumption $n \geq C_1 d_1' d_2'$. Therefore we have $\mathbb{E}_P[\Lambda_n^2] \leq C_6 b_{(2)}$. This further implies that

$$q_{1-\beta/2,n} \leq \ C_7 \max\left\{ \frac{\sqrt{2\mathbb{E}[\Lambda_n^2]}}{\beta^{1/2} n} \log\left(\frac{1}{\alpha}\right), \frac{1}{n^{3/2}} \log\left(\frac{1}{\alpha}\right) \right\}$$

$$\leq \frac{C_8}{\beta} \log\left(\frac{1}{\alpha}\right) \frac{b_{(2)}^{1/2}}{n},$$

where the last inequality uses $\beta \leq \beta^{1/2}$ and $n^{1/2} \geq C_1^{1/2} b_{(2)}^{-1/2}$ from the previous result (G.31). Hence the quantile is bounded as (G.30). This completes the proof of Lemma G.0.3. □

Let us come back to the proof of Proposition 8.9. Since type I error control is trivial by the union bound, we only need to show the type II error control of the adaptive test. As in the proof of Proposition 8.8, we know that there exists an integer $j \in \{1, \ldots, \gamma_{\max}^*\}$ such that

$$2^{j-1} < \left(\frac{n}{\log\log n}\right)^{\frac{2}{4s+d_1+d_2}} \leq 2^j. \tag{G.32}$$

We take such j and define $\kappa^* := 2^j \in \mathsf{K}^\dagger$. Since $\mathbb{P}_P(\phi^\dagger_{\text{adapt}} = 0) \leq \mathbb{P}_P(\phi^\dagger_{\kappa^*, \alpha/\gamma_{\max}^*} = 0)$, it suffices to show that the resulting multinomial test $\phi^\dagger_{\kappa^*, \alpha/\gamma_{\max}^*}$ controls the type II error by β under the given condition. To this end, we invoke Lemma G.0.3. Note that there are $(\kappa^*)^{d_1+d_2}$ number of bins for $\phi^\dagger_{\kappa^*, \alpha/\gamma_{\max}^*}$, which is bounded by

$$(\kappa^*)^{d_1+d_2} \overset{(i)}{\leq} 2^{d_1+d_2} \left(\frac{n}{\log\log n}\right)^{\frac{2(d_1+d_2)}{4s+d_1+d_2}} \overset{(ii)}{\leq} 2^{d_1+d_2} \left(\frac{n}{\log\log n}\right),$$

where (i) follows by the bound (G.32) and (ii) follows since $4s \geq d_1 + d_2$. Thus the condition of Lemma G.0.3 is fulfilled as the number of bins is smaller than the sample size n up to a constant factor which depends on d_1 and d_2. From the proof of Proposition 8.6, we know that

$$b_{(2)} = \max\{\|p_{YZ}\|_2^2, \|p_Y p_Z\|_2^2\} \leq L(\kappa^*)^{-(d_1+d_2)} \quad \text{and}$$

$$\|p_{YZ} - p_Y p_Z\|_2^2 \geq C_1(s, L, d_1, d_2)(\kappa^*)^{-(d_1+d_2)} \epsilon_n^2,$$

where ϵ_n is the lower bound for $\|f_{YZ} - f_Y f_Z\|_{L_2}$. Combining this observation with Lemma G.0.3 shows that $\phi^\dagger_{\kappa^*, \alpha/\gamma_{\max}^*}$ has non-trivial power when

$$\epsilon_n^2 \geq \frac{C_2(s, L, d_1, d_2)}{\beta} \log\left(\frac{\gamma_{\max}^*}{\alpha}\right) \frac{L^{1/2} \cdot (\kappa^*)^{(d_1+d_2)/2}}{n}.$$

By the definition of $\gamma_{\max}^*$ and κ^*, this inequality is further implied by

$$\epsilon_n^2 \geq C_3(s, L, d_1, d_2) \left(\frac{\log\log n}{n}\right)^{\frac{4s}{4s+d_1+d_2}}.$$

This completes the proof of Proposition 8.9.

G.19 Proof of Theorem 8.6

We use the quantile approach described in Section 8.3 to prove the result (see also Fromont et al., 2013). More specifically we let $q_{1-\beta/2,n}$ denote the quantile of the permutation critical value $c_{1-\alpha,n}$ of T_{χ^2}. Then as shown in the proof of Lemma 8.0.1, if

$$\mathbb{E}_P[T_{\chi^2}] \;\geq\; q_{1-\beta/2,n} + \sqrt{\frac{2\mathrm{Var}_P[T_{\chi^2}]}{\beta}}, \tag{G.33}$$

then the type II error of the permutation test is controlled as

$$\begin{aligned}\sup_{P\in\mathcal{P}_1} \mathbb{P}_P(T_{\chi^2} \leq c_{1-\alpha,n}) \;&\leq\; \sup_{P\in\mathcal{P}_1} \mathbb{P}_P(T_{\chi^2} \leq q_{1-\beta/2,n}) + \sup_{P\in\mathcal{P}_1} \mathbb{P}_P(q_{1-\beta/2,n} < c_{1-\alpha,n}) \\ &\leq\; \beta.\end{aligned}$$

Therefore we only need to show that the inequality (G.33) holds under the condition given in Theorem 8.6. Note that Chan et al. (2014) present a lower bound for $\mathbb{E}_P[T_{\chi^2}]$ as

$$\begin{aligned}\mathbb{E}_P[T_{\chi^2}] \;&=\; \sum_{k=1}^{d} \frac{\{p_Y(k)-p_Z(k)\}^2}{p_Y(k)+p_Z(k)} n\left(1 - \frac{1-e^{-n\{p_Y(k)+p_Z(k)\}}}{n\{p_Y(k)+p_Z(k)\}}\right) \\ &\geq\; \frac{n^2}{4d+2n}\|p_Y - p_Z\|_1^2,\end{aligned} \tag{G.34}$$

and an upper bound for $\mathrm{Var}_P[T_{\chi^2}]$ by

$$\mathrm{Var}_P[T_{\chi^2}] \;\leq\; 2\min\{n,d\} + 5n\sum_{k=1}^{d} \frac{\{p_Y(k)-p_Z(k)\}^2}{p_Y(k)+p_Z(k)}. \tag{G.35}$$

In the rest of the proof, we show that for some constant $C_1 > 0$,

$$q_{1-\beta/2,n} \leq \frac{C_1}{\beta}\log\left(\frac{1}{\alpha}\right)\sqrt{\min\{n,d\}}. \tag{G.36}$$

Building on these three observations (G.34), (G.35) and (G.36), we can verify that the sufficient condition (G.33) is satisfied under the assumption made in Theorem 8.6. Although it can be done by following Chan et al. (2014), their proof may be too concise for some readers (also there is a typo in their algorithm in Section 2 — the critical value should be $C\sqrt{\min\{n,d\}}$ instead of $C\sqrt{n}$) and so we decide to give detailed explanations in Appendix G.19.3. Hence all we need to show is condition (G.36).

G.19.1 Verification of condition (G.36)

Recall the permuted chi-square statistic $T^{\pi}_{\chi^2}$ given as

$$T^{\pi}_{\chi^2} = \sum_{k=1}^{d} \frac{(\sum_{i=1}^{n} X_{\pi_i,k} - \sum_{i=1}^{n} X_{\pi_{i+n},k})^2 - V_k - W_k}{V_k + W_k} \mathbb{1}(V_k + W_k > 0).$$

For simplicity, let us write $\omega_k := V_k + W_k$ for $k = 1, \ldots, d$. Note that $\omega_1, \ldots, \omega_d$ are permutation invariant and they should be constant under the permutation law. Having this observation in mind, we split the permuted statistic into two parts:

$$\begin{aligned} T^{\pi}_{\chi^2} &= \sum_{(i,j)\in \mathbf{i}_2^n} \sum_{k=1}^{d} \frac{(X_{\pi_i,k} - X_{\pi_{i+n},k})(X_{\pi_j,k} - X_{\pi_{j+n},k})}{\omega_k} \mathbb{1}(\omega_k > 0) \\ &\quad + \sum_{i=1}^{n} \sum_{k=1}^{d} \frac{(X_{\pi_i,k} - X_{\pi_{i+n},k})^2}{\omega_k} \mathbb{1}(\omega_k > 0) - \sum_{k=1}^{d} \mathbb{1}(\omega_k > 0) \\ &= T^{\pi}_{\chi^2,a} + T^{\pi}_{\chi^2,b} \quad \text{(say)}. \end{aligned}$$

Let us first compute an upper bound for the $1-\alpha$ critical value of $T^{\pi}_{\chi^2}$. To do so, recall that $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables. From the same reasoning made in Section 8.6.1, one can see that $T^{\pi}_{\xi^2,a}$ have the same distribution as

$$\sum_{(i,j)\in \mathbf{i}_2^n} \xi_i \xi_j \left[\sum_{k=1}^{d} \frac{(X_{\pi_i,k} - X_{\pi_{i+n},k})(X_{\pi_j,k} - X_{\pi_{j+n},k})}{\omega_k} \mathbb{1}(\omega_k > 0) \right].$$

Then following the same line of the proof of Theorem 8.3 with the trivial bound in (8.31), we have that for any $t > 0$,

$$\mathbb{P}_{\pi}\big(T^{\pi}_{\chi^2,a} \geq t \mid \mathcal{X}_n\big) \leq \exp\left\{ - C_2 \min\left(\frac{t^2}{\Sigma^2_{n,\text{pois}}}, \frac{t}{\Sigma_{n,\text{pois}}} \right) \right\}, \tag{G.37}$$

where

$$\Sigma^2_{n,\text{pois}} := \sum_{(i,j)\in \mathbf{i}_2^{2n}} \left\{ \sum_{k=1}^{d} \frac{X_{i,k} X_{j,k}}{\omega_k} \mathbb{1}(\omega_k > 0) \right\}^2 \tag{G.38}$$

and $\{X_{1,k}, \ldots, X_{2n,k}\} := \{Y_{1,k}, \ldots, Y_{n,k}, Z_{1,k}, \ldots, Z_{n,k}\}$. Also note that

$$\begin{aligned} T^{\pi}_{\chi^2,b} &= \sum_{i=1}^{n}\sum_{k=1}^{d} \frac{(X_{\pi_i,k} - X_{\pi_{i+n},k})^2}{\omega_k} \mathbb{1}(\omega_k > 0) - \sum_{k=1}^{d} \mathbb{1}(\omega_k > 0) \\ &\leq \sum_{i=1}^{2n}\sum_{k=1}^{d} \frac{X_{i,k}^2}{\omega_k} \mathbb{1}(\omega_k > 0) - \sum_{k=1}^{d} \mathbb{1}(\omega_k > 0) \\ &:= T_{\chi^2,b,\mathrm{up}} \end{aligned} \tag{G.39}$$

where $T_{\chi^2,b,\mathrm{up}}$ is independent of π. Furthermore, since each $X_{i,k}$ can have a nonnegative integer and $\omega_k = \sum_{i=1}^{2n} X_{i,k}$, it is clear that $\sum_{i=1}^{2n} X_{i,k}^2/\omega_k \geq 1$ whenever $\omega_k > 0$. This means that $T_{\chi^2,b,\mathrm{up}}$ is nonnegative. Combining the results (G.37) and (G.39), for any $t > 0$,

$$\mathbb{P}_\pi\big(T^{\pi}_{\chi^2} \geq t + T_{\chi^2,b,\mathrm{up}} \,|\mathcal{X}_n\big) \leq \mathbb{P}_\pi\big(T^{\pi}_{\chi^2,a} \geq t \,|\mathcal{X}_n\big) \leq \exp\left\{ - C_3 \min\left(\frac{t^2}{\Sigma^2_{n,\mathrm{pois}}}, \frac{t}{\Sigma_{n,\mathrm{pois}}}\right)\right\}.$$

By setting the upper bound to be α and assuming $\alpha < e^{-1}$, it can be seen that

$$c_{1-\alpha,n} \leq C_4 \Sigma_{n,\mathrm{pois}} \log\left(\frac{1}{\alpha}\right) + T_{\chi^2,b,\mathrm{up}}.$$

Let $q^*_{1-\beta/2,n}$ be the $1-\beta/2$ quantile of the above upper bound, which means that $q_{1-\beta/2,n} \leq q^*_{1-\beta/2,n}$. For now, we take the following two bounds for granted:

$$\mathbb{E}_P[\Sigma^2_{n,\mathrm{pois}}] \leq C_5 \min\{n, d\} \quad \text{and} \quad \mathbb{E}_P[T_{\chi^2,b,\mathrm{up}}] \leq C_6, \tag{G.40}$$

which are formally proved in Appendix G.19.2. Then by using Markov's inequality, for any $t_1, t_2 > 0$ and $t = t_1 + t_2$,

$$\begin{aligned} \mathbb{P}_P\big[C_5 \Sigma_{n,\mathrm{pois}} \log(\alpha^{-1}) + T_{\chi^2,b,\mathrm{up}} \geq t\big] &\leq \mathbb{P}_P\big[C_5 \Sigma_{n,\mathrm{pois}} \log(\alpha^{-1}) \geq t_1\big] + \mathbb{P}_P\big[T_{\chi^2,b,\mathrm{up}} \geq t_2\big] \\ &\leq C_6 \frac{\mathbb{E}_P[\Sigma^2_{n,\mathrm{pois}}]\{\log(\alpha^{-1})\}^2}{t_1^2} + C_7 \frac{\mathbb{E}_P[T_{\chi^2,b,\mathrm{up}}]}{t_2} \\ &\leq C_7 \frac{\min(n,d)\{\log(\alpha^{-1})\}^2}{t_1^2} + \frac{C_8}{t_2}. \end{aligned}$$

Then by setting the upper bound to be $\beta/2$, one may see that for sufficiently large $C_9 > 0$,

$$q^*_{1-\beta/2,n} \leq \frac{C_9}{\beta^{1/2}} \log\left(\frac{1}{\alpha}\right) \sqrt{\min\{n,d\}} + \frac{C_{10}}{\beta},$$

which in turn shows that condition (G.36) is satisfied.

G.19.2 Verification of two bounds in (G.40)

This section proves the bounds in (G.40), namely, (a) $\mathbb{E}_P[\Sigma^2_{n,\text{pois}}] \le C_1 \min\{n, d\}$ and (b) $\mathbb{E}_P[T_{\chi^2,b,\text{up}}] \le C_2$.

• **Bound (a).** We start by proving $\mathbb{E}_P[\Sigma^2_{n,\text{pois}}] \le C_1 \min\{n, d\}$. By recalling the definition of $\Sigma_{n,\text{pois}}$ in (G.38), note that

$$\begin{aligned}\Sigma^2_{n,\text{pois}} &= \sum_{(i,j)\in \mathbf{i}^n_2} \left\{ \sum_{k=1}^d \frac{Y_{i,k}Y_{j,k}}{\omega_k} \mathbb{1}(\omega_k > 0) \right\}^2 + \sum_{(i,j)\in \mathbf{i}^n_2} \left\{ \sum_{k=1}^d \frac{Z_{i,k}Z_{j,k}}{\omega_k} \mathbb{1}(\omega_k > 0) \right\}^2 \\ &\quad + 2 \sum_{1\le i,j\le n} \left\{ \sum_{k=1}^d \frac{Y_{i,k}Z_{j,k}}{\omega_k} \mathbb{1}(\omega_k > 0) \right\}^2 \\ &:= \Sigma^2_{n,Y} + \Sigma^2_{n,Z} + 2\Sigma^2_{n,YZ} \quad \text{(say)}.\end{aligned}$$

Given $1 \le i \ne j \le n$, expand the first squared term as

$$\begin{aligned}\left\{ \sum_{k=1}^d \frac{Y_{i,k}Y_{j,k}}{\omega_k} \mathbb{1}(\omega_k > 0) \right\}^2 &= \sum_{k=1}^d \omega_k^{-2} Y^2_{i,k} Y^2_{j,k} \mathbb{1}(\omega_k > 0) \\ &\quad + \sum_{(k_1,k_2)\in \mathbf{i}^d_2} \omega_{k_1}^{-1}\omega_{k_2}^{-1} Y_{i,k_1}Y_{j,k_1}Y_{i,k_2}Y_{j,k_2} \mathbb{1}(\omega_{k_1} > 0)\mathbb{1}(\omega_{k_2} > 0) \\ &= (I) + (II) \quad \text{(say)}.\end{aligned}$$

Let us first look at the expectation of (I). Suppose that $Q_1, \ldots, Q_n$ are independent Poisson random variables with parameters $\lambda_1, \ldots, \lambda_n$, respectively. To calculate the above expectation, we use the fact that conditional on the event $\sum_{i=1}^n Q_i = N$, $(Q_1, \ldots, Q_n)$ has a multinomial distribution as

$$(Q_1, \ldots, Q_n) \sim \text{Multinomial}\left(N, \left\{ \frac{\lambda_1}{\sum_{i=1}^n \lambda_i}, \ldots, \frac{\lambda_n}{\sum_{i=1}^n \lambda_i} \right\} \right).$$

Therefore, conditioned on $\omega_k = N$, we observe that

$$\begin{aligned}&(Y_{i,k}, Y_{j,k}, \omega_k - Y_{i,k} - Y_{j,k}) \\ \sim\ &\text{Multinomial}\left(N, \left[\frac{p_Y(k)}{n\{p_Y(k)+p_Z(k)\}}, \frac{p_Y(k)}{n\{p_Y(k)+p_Z(k)\}}, 1 - \frac{2p_Y(k)}{n\{p_Y(k)+p_Z(k)\}} \right] \right).\end{aligned}$$

Using this property and the moment generating function (MGF) of a multinomial distribution (see Appendix G.19.4),

$$\begin{aligned}\mathbb{E}\big[Y_{i,k}^2 Y_{j,k}^2 \mid \omega_k = N\big] &= N(N-1)(N-2)(N-3)\widetilde{p}_{k,n}^4 \\ &\quad + 2N(N-1)(N-2)\widetilde{p}_{k,n}^3 + N(N-1)\widetilde{p}_{k,n}^2,\end{aligned}$$

where

$$\widetilde{p}_{k,n} := \frac{p_Y(k)}{n\{p_Y(k)+p_Z(k)\}}.$$

This gives

$$\begin{aligned}\mathbb{E}\left[\frac{Y_{i,k}^2 Y_{j,k}^2}{\omega_k^2}\mathbb{1}(\omega_k>0)\right] &= \mathbb{E}_N\left[\frac{Y_{i,k}^2 Y_{j,k}^2}{\omega_k^2}\mathbb{1}(\omega_k>0)\;\middle|\;\omega_k = N\right\}\Bigg] \\ &\le \mathbb{E}_N\left[N^2\widetilde{p}_{k,n}^4\mathbb{1}(N>0)\right] + 2\mathbb{E}_N\left[N\widetilde{p}_{k,n}^3\mathbb{1}(N>0)\right] + \mathbb{E}_N\left[\widetilde{p}_{k,n}^2\mathbb{1}(N>0)\right].\end{aligned}$$

By noting that $N \sim \text{Poisson}(n\{p_Y(k)+p_Z(k)\})$,

$$\begin{aligned}\mathbb{E}_N\left[N^2\widetilde{p}_{k,n}^4\mathbb{1}(N>0)\right] &= \widetilde{p}_{k,n}^4\mathbb{E}_N\left[N^2\mathbb{1}(N>0)\right] \\ &= \left(\frac{p_Y(k)}{n\{p_Y(k)+p_Z(k)\}}\right)^4 (n\{p_Y(k)+p_Z(k)\})^2 \\ &\quad + \left(\frac{p_Y(k)}{n\{p_Y(k)+p_Z(k)\}}\right)^4 n\{p_Y(k)+p_Z(k)\} \\ &\le \frac{p_Y(k)^4}{(n\{p_Y(k)+p_Z(k)\})^2} + \frac{p_Y(k)^4}{(n\{p_Y(k)+p_Z(k)\})^3},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_N\left[N\widetilde{p}_{k,n}^3\mathbb{1}(N>0)\right] &= \widetilde{p}_{k,n}^3\mathbb{E}_N\left[N\mathbb{1}(N>0)\right] \\ &= \frac{p_Y(k)^3}{(n\{p_Y(k)+p_Z(k)\})^2}, \\ \mathbb{E}_N\left[\widetilde{p}_{k,n}^2\mathbb{1}(N>0)\right] &= \widetilde{p}_{k,n}^2\mathbb{E}_N\left[\mathbb{1}(N>0)\right] \\ &= \left(\frac{p_Y(k)}{n\{p_Y(k)+p_Z(k)\}}\right)^2 \times \left(1-e^{-n\{p_Y(k)+p_Z(k)\}}\right).\end{aligned}$$

Putting these together,

$$\begin{aligned}
\mathbb{E}[(I)] &= \sum_{k=1}^{d} \mathbb{E}\left[\frac{Y_{i,k}^2 Y_{j,k}^2}{\omega_k^2} \mathbb{1}(\omega_k > 0)\right] \\
&\leq \sum_{k=1}^{d} \frac{p_Y(k)^4}{(n\{p_Y(k)+p_Z(k)\})^2} + \sum_{k=1}^{d} \frac{p_Y(k)^4}{(n\{p_Y(k)+p_Z(k)\})^3} + \sum_{k=1}^{d} \frac{p_Y(k)^3}{(n\{p_Y(k)+p_Z(k)\})^2} \\
&\quad + \sum_{k=1}^{d} \left(\frac{p_Y(k)}{n\{p_Y(k)+p_Z(k)\}}\right)^2 \times \left(1 - e^{-n\{p_Y(k)+p_Z(k)\}}\right) \\
&\leq \frac{1}{n^2} + \frac{1}{n^3} + \frac{1}{n^2} + \frac{1}{n^2} \min\{d, 2n\}, \qquad \text{(G.41)}
\end{aligned}$$

where the last inequality uses $1 - e^{-x} \leq \min\{x, 1\}$.

Next moving onto the expected value of (II), the independence between Poisson random variables gives

$$\begin{aligned}
&\mathbb{E}\left[\frac{Y_{i,k_1} Y_{j,k_1}}{\omega_{k_1}} \frac{Y_{i,k_2} Y_{j,k_2}}{\omega_{k_2}} \mathbb{1}(\omega_{k_1} > 0) \mathbb{1}(\omega_{k_2} > 0)\right] \\
= \; &\mathbb{E}\left[\frac{Y_{i,k_1} Y_{j,k_1}}{\omega_{k_1}} \mathbb{1}(\omega_{k_2} > 0)\right] \mathbb{E}\left[\frac{Y_{i,k_2} Y_{j,k_2}}{\omega_{k_2}} \mathbb{1}(\omega_{k_1} > 0)\right].
\end{aligned}$$

Again, $(Y_{i,k_1}, Y_{j,k_1}, \omega_{k_1} - Y_{i,k_1} - Y_{j,k_1})$ has a multinomial distribution conditional on $\omega_{k_1} = N$. Based on this property, we have

$$\mathbb{E}[Y_{i,k_1} Y_{j,k_1} | \omega_{k_1} = N] = N(N-1)\widetilde{p}_{k_1,n}^2.$$

Thus

$$\begin{aligned}
\mathbb{E}\left[\frac{Y_{i,k_1} Y_{j,k_1}}{\omega_{k_1}} \mathbb{1}(\omega_{k_1} > 0)\right] &= \mathbb{E}_N\left[\mathbb{E}\left\{\frac{Y_{i,k_1} Y_{j,k_1}}{\omega_{k_1}} \mathbb{1}(\omega_{k_1} > 0) \middle| \omega_{k_1} = N\right\}\right] \\
&= \widetilde{p}_{k_1,n}^2 \mathbb{E}_N\left[(N-1)\mathbb{1}(N > 0)\right] \\
&= \widetilde{p}_{k_1,n}^2 \left[n\{p_Y(k_1) + p_Z(k_1)\} - 1 + e^{-n\{p_Y(k_1)+p_Z(k_1)\}}\right] \\
&\leq \frac{p_Y^2(k_1)}{n\{p_Y(k_1) + p_Z(k_1)\}}.
\end{aligned}$$

This gives

$$\mathbb{E}[(II)] = \mathbb{E}\left[\sum_{(k_1,k_2) \in \mathrm{i}_2^d} \omega_{k_1}^{-1} \omega_{k_2}^{-1} Y_{i,k_1} Y_{j,k_1} Y_{i,k_2} Y_{j,k_2} \mathbb{1}(\omega_{k_1} > 0) \mathbb{1}(\omega_{k_2} > 0)\right]$$

$$\leq \left(\sum_{i_1=1}^{d} \frac{p_{i_1}^2}{n\{p_{i_1}+q_{i_1}\}}\right) \cdot \left(\sum_{i_2=1}^{d} \frac{p_{i_2}^2}{n\{p_{i_2}+q_{i_2}\}}\right) \leq \frac{1}{n^2}. \tag{G.42}$$

Therefore based on (G.41) and (G.42), it is clear that $\mathbb{E}_P[\Sigma_{n,Y}^2] \leq C_2 \min\{n,d\}$. The same analysis further shows that $\mathbb{E}_P[\Sigma_{n,Z}^2] \leq C_3 \min\{n,d\}$ and $\mathbb{E}_P[\Sigma_{n,YZ}^2] \leq C_4 \min\{n,d\}$, which leads to $\mathbb{E}_P[\Sigma_{n,\text{pois}}^2] \leq C_1 \min\{n,d\}$ as desired.

• **Bound (b).** Next we prove that $\mathbb{E}_P[T_{\chi^2,b,\text{up}}] \leq C_2$. Recall that $T_{\chi^2,b,\text{up}}$ is a nonnegative random variable defined in (G.39). Since $\omega_k \sim \text{Poisson}(n\{p_Y(k)+p_Z(k)\})$, the second term of $T_{\chi^2,b,\text{up}}$ satisfies

$$\sum_{k=1}^{d} \mathbb{E}[\mathbb{1}(\omega_k > 0)] = \sum_{k=1}^{d} \left(1 - e^{-n\{p_Y(k)+p_Z(k)\}}\right).$$

Next consider the first term of $T_{\chi^2,b,\text{up}}$:

$$\sum_{i=1}^{2n}\sum_{k=1}^{d} \frac{X_{i,k}^2}{\omega_k}\mathbb{1}(\omega_k > 0).$$

Note that based on the moments of a multinomial distribution (see Appendix G.19.4), one can compute

$$\mathbb{E}\big[Y_{i,k}^2 | \omega_k = N\big] = N(N-1)\widetilde{p}_{k,n}^2 + n\widetilde{p}_{k,n},$$
$$\mathbb{E}\big[Z_{i,k}^2 | \omega_k = N\big] = N(N-1)\widetilde{q}_{k,n}^2 + n\widetilde{q}_{k,n},$$

where $\widetilde{p}_{k,n} := p_Y(k)/\{n(p_Y(k)+p_Z(k)\}$ and $\widetilde{q}_{k,n} := p_Z(k)/\{n(p_Y(k)+p_Z(k))\}$. Therefore, by the law of total expectation,

$$\begin{aligned}
\mathbb{E}\left[\frac{Y_{i,k}^2}{\omega_k}\mathbb{1}(\omega_k > 0)\right] &= \mathbb{E}_N\left[\mathbb{E}\left\{\frac{Y_{i,k}^2}{\omega_k}\mathbb{1}(\omega_k > 0)\middle|\omega_k = N\right\}\right] \\
&= \widetilde{p}_{k,n}^2 \mathbb{E}_N\left[(N-1)\mathbb{1}(N>0)\right] + \widetilde{p}_{k,n}\mathbb{E}_N\left[\mathbb{1}(N>0)\right] \\
&= \widetilde{p}_{k,n}^2\left(n\{p_Y(k)+p_Z(k)\} - 1 + e^{-n\{p_Y(k)+p_Z(k)\}}\right) \\
&\quad + \widetilde{p}_{k,n}\left(1 - e^{-n\{p_Y(k)+p_Z(k)\}}\right).
\end{aligned}$$

Similarly, one can compute

$$\mathbb{E}\left[\frac{Z_{i,k}^2}{\omega_k}\mathbb{1}(\omega_k > 0)\right] = \widetilde{q}_{k,n}^2\left(n\{p_Y(k)+p_Z(k)\} - 1 + e^{-n\{p_Y(k)+p_Z(k)\}}\right)$$

$$+\widetilde{q}_{k,n}\left(1-e^{-n\{p_Y(k)+p_Z(k)\}}\right).$$

Based on the definition of $\widetilde{p}_{k,n}$ and $\widetilde{q}_{k,n}$, we have the identity

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{k=1}^{d}\widetilde{p}_{k,n}\left(1-e^{-n\{p_Y(k)+p_Z(k)\}}\right)+\sum_{i=1}^{n}\sum_{k=1}^{d}\widetilde{q}_{k,n}\left(1-e^{-n\{p_Y(k)+p_Z(k)\}}\right)\\
=\;&\sum_{k=1}^{d}\left(1-e^{-n\{p_Y(k)+p_Z(k)\}}\right),
\end{aligned}$$

which is the expected value of $\sum_{k=1}^{d}\mathbb{1}(\omega_k>0)$. Putting everything together,

$$\begin{aligned}
\mathbb{E}\big[T_{\chi^2,b,\mathrm{up}}\big] &= \sum_{i=1}^{2n}\sum_{k=1}^{d}\mathbb{E}\left[\frac{X_{i,k}^2}{\omega_k}\mathbb{1}(\omega_k>0)\right]-\sum_{k=1}^{d}\mathbb{E}[\mathbb{1}(\omega_k>0)]\\
&\le \sum_{i=1}^{n}\sum_{k=1}^{d}\frac{p_Y^2(k)}{n\{p_Y(k)+p_Z(k)\}}+\sum_{i=1}^{n}\sum_{k=1}^{d}\frac{p_Z^2(k)}{n\{p_Y(k)+p_Z(k)\}}\\
&\le 2.
\end{aligned}$$

This proves the bound $\mathbb{E}[T_{\chi^2,b,\mathrm{up}}]\le C_2$.

G.19.3 Details on verifying the sufficient condition (G.33)

First assume that $n<d$. Then the variance (G.35) is dominated by the first term and thus condition (G.33) is fulfilled when

$$\begin{aligned}
\mathbb{E}_P[T_{\chi^2}] &\overset{(i)}{\ge} \frac{n^2}{6d}\|p_Y-p_Z\|_1^2 \overset{(ii)}{\ge} \frac{n^2}{6d}\epsilon_n^2 \overset{(iii)}{\ge} \frac{C_2}{\beta^{1/2}}\log\left(\frac{1}{\alpha}\right)\sqrt{n}\\
&\ge q_{1-\beta/2,n}+\sqrt{\frac{2\mathrm{Var}_P[T_{\chi^2}]}{\beta}},
\end{aligned}$$

where (i) follows by the bound (G.34), (ii) uses $\|p_Y-p_Z\|_1\ge\epsilon_n$ and (iii) holds from the bounds (G.35) and (G.36) and the condition on ϵ_n, i.e.

$$\epsilon_n \;\ge\; \frac{C_3}{\beta^{1/2}}\sqrt{\log\left(\frac{1}{\alpha}\right)}\frac{d^{1/2}}{n^{3/4}},$$

for some large constant $C_3>0$.

Next assume that $n \geq d$. For convenience, let us write

$$\varphi_k := 1 - \frac{1 - e^{-np_Y(k) - np_Z(k)}}{np_Y(k) + np_Z(k)} \quad \text{for } k = 1, \ldots, d.$$

We define $\mathsf{I}_d := \{k \in \{1, \ldots, d\} : 2\varphi_k \geq 1\}$ and denote its complement by I_d^c. Note that $np_Y(k) + np_Z(k) < 2$ for $k \in \mathsf{I}_d^c$ and thus

$$\begin{aligned} n \sum_{k=1}^{d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)} &= n \sum_{k \in \mathsf{I}_d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)} + n \sum_{k \in \mathsf{I}_d^c} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)} \\ &\leq n \sum_{k \in \mathsf{I}_d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)} + 2d. \end{aligned}$$

Based on this observation with $n \geq d$, the variance of T_{χ^2} can be further bounded by

$$\mathrm{Var}_P[T_{\chi^2}] \leq 4d + 5n \sum_{k \in \mathsf{I}_d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)}. \tag{G.43}$$

Let us make one more observation that $n^2 \|p_Y - p_Z\|_1^2 / (4d + 2n) \geq C_4 \beta^{-1}$ for some large constant $C_4 > 0$, which holds under the assumption on ϵ_n in Theorem 8.6 and $n \geq d$. Based on this, the expectation of T_{χ^2} is bounded by

$$\begin{aligned} \mathbb{E}_P[T_{\chi^2}] &\geq \sqrt{\sum_{k=1}^{d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)} n\varphi_k} \sqrt{\frac{n^2}{4d + 2n} \|p_Y - p_Z\|_1^2} \\ &\geq \sqrt{\frac{C_4 n}{2\beta} \sum_{k \in \mathsf{I}_d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)}}, \end{aligned} \tag{G.44}$$

where the last inequality uses the definition of I_d. This gives

$$\begin{aligned} \mathbb{E}_P[T_{\chi^2}] &\overset{(i)}{\geq} \frac{1}{2}\mathbb{E}_P[T_{\chi^2}] + \frac{1}{2}\sqrt{\frac{C_4 n}{2\beta} \sum_{k \in \mathsf{I}_d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)}} \\ &\overset{(ii)}{\geq} \frac{n}{12}\|p_Y - p_Z\|_1^2 + \frac{1}{2}\sqrt{\frac{C_4 n}{2\beta} \sum_{k \in \mathsf{I}_d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)}} \\ &\overset{(iii)}{\geq} \frac{n}{12}\epsilon_n^2 + \frac{1}{2}\sqrt{\frac{C_4 n}{2\beta} \sum_{k \in \mathsf{I}_d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)}} \\ &\overset{(iv)}{\geq} \frac{C_5}{\beta} \log\left(\frac{1}{\alpha}\right) d^{1/2} + \frac{1}{2}\sqrt{\frac{C_4 n}{2\beta} \sum_{k \in \mathsf{I}_d} \frac{\{p_Y(k) - p_Z(k)\}^2}{p_Y(k) + p_Z(k)}} \end{aligned}$$

$$\overset{(v)}{\geq} \frac{C_1}{\beta}\log\left(\frac{1}{\alpha}\right)d^{1/2} + \sqrt{\frac{8d}{\beta}} + \sqrt{\frac{10n}{\beta}\sum_{k\in \mathsf{I}_d}\frac{\{p_Y(k)-p_Z(k)\}^2}{p_Y(k)+p_Z(k)}}$$

$$\overset{(vi)}{\geq} q_{1-\beta/2,n} + \sqrt{\frac{2\mathrm{Var}_P[T_{\chi^2}]}{\beta}},$$

where (i) uses the lower bound (G.44), (ii) and (iii) follow by the bound (G.34) and $\|p_Y - p_Z\|_1 \geq \epsilon_n$, respectively, (iv) follows from the lower bound for ϵ_n in the theorem statement, (v) holds by choosing C_4, C_5 large and lastly (vi) uses (G.36) and (G.43).

G.19.4 Multinomial Moments

This section collects some moments of a multinomial distribution that are used in the proof of Theorem 8.6. Suppose that $\mathbf{X} = (X_1, \ldots, X_d)$ has a multinomial distribution with the number of trials n and probabilities $(p_1, \ldots, p_d)$. The MGF of $\mathbf{X}$ is given by

$$M_{\mathbf{X}}(\mathbf{t}) = \left(\sum_{i=1}^{d} p_i e^{t_i}\right)^n.$$

We collect some of partial derivatives of the MGF.

$$\begin{aligned}
\frac{\partial}{\partial t_i}M_{\mathbf{X}}(\mathbf{t}) &= n\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-1} p_i e^{t_i},\\
\frac{\partial^2}{\partial t_i \partial t_j}M_{\mathbf{X}}(\mathbf{t}) &= n(n-1)\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-2} p_i e^{t_i} p_j e^{t_j},\\
\frac{\partial^2}{\partial t_i^2}M_{\mathbf{X}}(\mathbf{t}) &= n(n-1)\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-2} p_i^2 e^{2t_i} + n\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-1} p_i e^{t_i},\\
\frac{\partial^3}{\partial t_i^2 \partial t_j}M_{\mathbf{X}}(\mathbf{t}) &= n(n-1)(n-2)\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-3} p_i^2 p_j e^{2t_i} e^{t_j}\\
&\quad + n(n-1)\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-2} p_i p_j e^{t_i} e^{t_j},\\
\frac{\partial^4}{\partial t_i^2 \partial t_j^2}M_{\mathbf{X}}(\mathbf{t}) &= n(n-1)(n-2)(n-3)\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-4} p_i^2 p_j^2 e^{2t_i} e^{2t_j}\\
&\quad + n(n-1)(n-2)\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-3} p_i^2 p_j e^{2t_i} e^{t_j}
\end{aligned}$$

$$+ n(n-1)(n-2)\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-3} p_i p_j^2 e^{t_i} e^{2t_j}$$
$$+ n(n-1)\left(\sum_{i=1}^{d} p_i e^{t_i}\right)^{n-2} p_i p_j e^{t_i} e^{t_j}.$$

By setting $\mathbf{t} = 0$, for $i \neq j$,

$$\begin{aligned}
\mathbb{E}[X_i] &= np_i, \\
\mathbb{E}[X_i^2] &= n(n-1)p_i^2 + np_i, \\
\mathbb{E}[X_i X_j] &= n(n-1)p_i p_j, \\
\mathbb{E}[X_i^2 X_j] &= n(n-1)(n-2)p_i^2 p_j + n(n-1)p_i p_j, \\
\mathbb{E}[X_i^2 X_j^2] &= n(n-1)(n-2)(n-3)p_i^2 p_j^2 + n(n-1)(n-2)p_i^2 p_j \\
&\quad + n(n-1)(n-2)p_i p_j^2 + n(n-1)p_i p_j.
\end{aligned}$$

G.20 Proof of Proposition 8.10

Recall that the test is carried out via sample-splitting and the critical value of the permutation test is obtained by permuting the labels within $\mathcal{X}_{2n_1}^{\text{split}} = \{Y_1, \ldots, Y_{n_1}, Z_1, \ldots, Z_{n_1}\}$. Nevertheless, the distribution of the test statistic is invariant to any partial permutation under the null hypothesis. Based on this property, it can be shown that type I error control of the permutation test via sample-splitting is also guaranteed (see e.g. Theorem 15.2.1 of Lehmann and Romano, 2006). Hence we focus on the type II error control. Note that conditional on $w_1, \ldots, w_d$, the test statistic $U_{n_1,n_2}^{\text{split}}$ can be viewed as a U-statistic with kernel $g_{\text{Multi},w}(x, y)$ given in (8.42). Moreover this U-statistic is based on the two samples of equal size, which allows us to apply Lemma G.0.1. Based on this observation, we first study the performance of the test conditioning on $w_1, \ldots, w_d$. We then remove this conditioning part using Markov's inequality and conclude the result.

• **Conditional Analysis.** In this part, we investigate the type II error of the permutation test conditional on $w_1, \ldots, w_d$. As noted earlier, $U_{n_1,n_2}^{\text{split}}$ can be viewed as a U-statistic and so we can apply Lemma G.0.1 to proceed. To do so, we need to lower bound the conditional expectation of $U_{n_1,n_2}^{\text{split}}$ and upper bound $\psi_{Y,1}(P)$, $\psi_{Z,1}(P)$ and $\psi_{YZ,2}(P)$. On the one hand, the conditional expectation of $U_{n_1,n_2}^{\text{split}}$ is lower bounded by the squared ℓ_1 distance as

$$\mathbb{E}_P\big[U_{n_1,n_2}^{\text{split}} | w_1, \ldots, w_n\big] = \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k} \geq \|p_Y - p_Z\|_1^2,$$

where the inequality follows by Cauchy-Schwarz inequality and $\sum_{k=1}^{d} w_k = 1$. On the other hand, $\psi_{Y,1}(P)$, $\psi_{Z,1}(P)$ and $\psi_{YZ,2}(P)$ are upper bounded by

$$\begin{aligned}
\psi_{Y,1}(P) &\leq 4\sqrt{\sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k^2}} \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k} \\
\psi_{Z,1}(P) &\leq 4\sqrt{\sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2}} \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k} \\
\psi_{YZ,2}(P) &\leq \max\left\{ \sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k^2}, \sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2} \right\}.
\end{aligned} \tag{G.45}$$

The details of the derivations are presented in Section G.20.1. Further note that

$$\begin{aligned}
\sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k^2} &\overset{(i)}{\leq} 2\sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k^2} + 2\sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2} \\
&\overset{(ii)}{\leq} 4d\sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k} + 2\sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2},
\end{aligned} \tag{G.46}$$

where (i) uses $(x+y)^2 \leq 2x^2 + 2y^2$ and (ii) follows since $w_k \geq 1/(2d)$ for $k = 1, \ldots, d$. For notational convenience, let us write

$$\begin{aligned}
\|p_Y - p_Z\|_w^2 &:= \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k} \quad \text{and} \\
\|p_Z/w\|_2^2 &:= \sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2}.
\end{aligned}$$

Having this notation in place, we see that the condition (G.8) in Lemma G.0.1 is fulfilled when

$$\begin{aligned}
&\sqrt{\frac{\psi_{Y,1}(P)}{\beta n_1}} \leq C_1 \|p_Y - p_Z\|_w^2, \\
&\sqrt{\frac{\psi_{Z,1}(P)}{\beta n_1}} \leq C_2 \|p_Y - p_Z\|_w^2 \quad \text{and} \\
&\sqrt{\frac{\psi_{YZ,2}(P)}{\beta}} \log\left(\frac{1}{\alpha}\right) \frac{1}{n_1} \leq C_3 \|p_Y - p_Z\|_w^2.
\end{aligned}$$

Based on the results in (G.45) and (G.46), it can be shown that these three inequalities are implied by

$$\|p_Y - p_Z\|_w^2 \;\geq\; \frac{C_4}{\beta} \log\left(\frac{1}{\alpha}\right) \frac{\|p_Z/w\|_2}{n_1} \quad \text{and}$$
$$\|p_Y - p_Z\|_w^2 \;\geq\; \frac{C_5}{\beta^2} \log^2\left(\frac{1}{\alpha}\right) \frac{d}{n_1^2}.$$

Moreover, using the lower bound of the conditional expectation $\|p_Y - p_Z\|_w^2 \geq \|p_Y - p_Z\|_1^2 \geq \epsilon_{n_1,n_2}^2$ and the boundedness of ℓ_1 norm so that $\epsilon_{n_1,n_2}^2 \leq 4$, the above two inequalities are further implied by

$$\epsilon_{n_1,n_2}^2 \;\geq\; \frac{C_4}{\beta} \log\left(\frac{1}{\alpha}\right) \frac{\|p_Z/w\|_2}{n_1} \quad \text{and}$$
$$\epsilon_{n_1,n_2}^2 \;\geq\; \frac{2\sqrt{C_5}}{\beta} \log\left(\frac{1}{\alpha}\right) \frac{d^{1/2}}{n_1}.$$

In other words, for a sufficiently large $C_6 > 0$, the type II error of the permutation test is at most β when

$$\epsilon_{n_1,n_2}^2 \geq \frac{C_6}{\beta} \log\left(\frac{1}{\alpha}\right) \max\left\{\frac{\|p_Z/w\|_2}{n_1},\; \frac{d^{1/2}}{n_1}\right\}. \tag{G.47}$$

Note that the above condition is not deterministic as $w_1, \ldots, w_d$ are random variables. Next we remove this randomness.

• **Unconditioning $w_1, \ldots, w_d$.** Recall that $m = \min\{n_2, d\}$ and thus w_k is clearly lower bounded by

$$w_k = \frac{1}{2d} + \frac{1}{2m} \sum_{i=1}^{m} \mathbb{1}(Z_{i+n_2} = k) \geq \frac{1}{2d}\left[1 + \sum_{i=1}^{m} \mathbb{1}(Z_{i+n_2} = k)\right].$$

Based on this bound, one can see that $\|p_Z/w\|_2^2$ has the expected value upper bounded by

$$\begin{aligned}
\mathbb{E}_P\left[\sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2}\right] \;&\leq\; 4d^2 \sum_{k=1}^{d} \mathbb{E}_P\left[\frac{p_Z^2(k)}{\{1 + \sum_{i=1}^{m} \mathbb{1}(Z_{i+n_2} = k)\}^2}\right] \\
&\leq\; 4d^2 \sum_{k=1}^{d} \mathbb{E}_P\left[\frac{p_Z^2(k)}{1 + \sum_{i=1}^{m} \mathbb{1}(Z_{i+n_2} = k)}\right] \\
&\overset{(i)}{\leq}\; 4d^2 \sum_{k=1}^{d} \frac{p_Z^2(k)}{(m+1)p_Z(k)} \\
&\leq\; \frac{4d^2}{m},
\end{aligned}$$

where (i) uses the fact that when $X \sim \text{Binominal}(n, p)$, we have

$$\mathbb{E}\left[\frac{1}{1+X}\right] = \frac{1-(1-p)^{n+1}}{(n+1)p} \leq \frac{1}{(n+1)p}. \tag{G.48}$$

See e.g. Canonne et al. (2018) for the proof. Using this upper bound of the expected value, Markov's inequality yields

$$\mathbb{P}_P\left(\sqrt{\sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2}} \geq t\right) \leq \frac{1}{t^2}\mathbb{E}\left[\sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2}\right] \leq \frac{4d^2}{mt^2}.$$

By letting the right-hand side be β and $\mathcal{A}$ be the event such that $\mathcal{A} := \{\|p_Z/w\|_2 < 2d/\sqrt{m\beta}\}$, we know that $\mathbb{P}_P(\mathcal{A}) \geq 1-\beta$. Under this good event $\mathcal{A}$, the sufficient condition (G.47) is fulfilled when

$$\begin{aligned} \epsilon_{n_1,n_2}^2 &\geq \frac{C_7}{\beta^{3/2}} \log\left(\frac{1}{\alpha}\right) \max\left\{\frac{d}{\sqrt{m}n_1}, \frac{d^{1/2}}{n_1}\right\} \\ &= \frac{C_7}{\beta^{3/2}} \log\left(\frac{1}{\alpha}\right) \max\left\{\frac{d}{n_1\sqrt{n_2}}, \frac{d^{1/2}}{n_1}\right\}. \end{aligned} \tag{G.49}$$

• **Completion of the proof.** To complete the proof, let us denote the critical value of the permutation test by $c_{1-\alpha,n_1,n_2}$. Then the type II error of the permutation test is bounded by

$$\begin{aligned} \mathbb{P}_P(U_{n_1,n_2}^{\text{split}} \leq c_{1-\alpha,n_1,n_2}) &= \mathbb{P}_P(U_{n_1,n_2}^{\text{split}} \leq c_{1-\alpha,n_1,n_2}, \mathcal{A}) + \mathbb{P}_P(U_{n_1,n_2}^{\text{split}} \leq c_{1-\alpha,n_1,n_2}, \mathcal{A}^c) \\ &\leq \mathbb{P}_P(U_{n_1,n_2}^{\text{split}} \leq c_{1-\alpha,n_1,n_2}, \mathcal{A}) + \mathbb{P}_P(\mathcal{A}^c). \end{aligned}$$

As shown before, the type II error under the event $\mathcal{A}$ is bounded by β, which leads to $\mathbb{P}_P(U_{n_1,n_2}^{\text{split}} \leq c_{1-\alpha,n_1,n_2}, \mathcal{A}) \leq \beta$. Also we have $\mathbb{P}_P(\mathcal{A}^c) \leq \beta$ proved by Markov's inequality. Thus the unconditional type II error is bounded by 2β. Notice that condition (G.49) is equivalent to condition (8.43) given in Proposition 8.10. Hence the proof is completed by letting $2\beta = \beta'$.

G.20.1 Details on Equation (G.45)

We start with bounding $\psi_{Y,1}(P)$. Following the proof of Proposition 8.1, it can be seen that

$$\begin{aligned} \psi_{Y,1}(P) &= \mathbb{E}_P\left[\left(\sum_{k=1}^{d} w_k^{-1}[\mathbb{1}(Y_1 = k) - p_Y(k)][p_Y(k) - p_Z(k)]\right)^2 \middle| w_1, \ldots, w_d\right] \\ &\leq 2\sum_{k=1}^{d} w_k^{-2} p_Y(k)[p_Y(k) - p_Z(k)]^2 + 2\left(\sum_{k=1}^{d} w_k^{-1} p_Y(k)[p_Y(k) - p_Z(k)]\right)^2 \end{aligned}$$

$$:= \; 2(I) + 2(II).$$

For the first term (I), we apply Cauchy-Schwarz inequality to have

$$\begin{aligned}\sum_{k=1}^{d} w_k^{-2} p_Y(k)[p_Y(k) - p_Z(k)]^2 \leq & \sqrt{\sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k^2}} \sqrt{\sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^4}{w_k^2}} \\ \leq & \sqrt{\sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k^2}} \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k},\end{aligned}$$

where the second inequality follows by the monotonicity of ℓ_p norm. For the second term (II), we apply Cauchy-Schwarz inequality repeatedly to have

$$\begin{aligned}\left(\sum_{k=1}^{d} w_k^{-1} p_Y(k)[p_Y(k) - p_Z(k)] \right)^2 \leq & \sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k} \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k} \\ \leq & \sqrt{\sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k^2}} \sqrt{\sum_{k=1}^{d} p_Y^2(k)} \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k} \\ \overset{(i)}{\leq} & \sqrt{\sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k^2}} \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k}\end{aligned}$$

where (i) uses $\sum_{k=1}^{d} p_Y^2(k) \leq 1$. Combining the results yields

$$\psi_{Y,1}(P) \leq 4 \sqrt{\sum_{k=1}^{d} \frac{p_Y^2(k)}{w_k^2}} \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k}.$$

By symmetry, it similarly follows that

$$\psi_{Z,1}(P) \leq 4 \sqrt{\sum_{k=1}^{d} \frac{p_Z^2(k)}{w_k^2}} \sum_{k=1}^{d} \frac{[p_Y(k) - p_Z(k)]^2}{w_k}.$$

These establish the first two inequalities in (G.45). Next we find an upper bound for $\psi_{YZ,2}(P)$. By recalling the definition of $\psi_{YZ,2}(P)$, we have

$$\begin{aligned}\psi_{YZ,2}(P) := \max \big\{ & \mathbb{E}_P[g^2_{\text{Multi},w}(Y_1, Y_2) | w_1, \ldots, w_d], \; \mathbb{E}_P[g^2_{\text{Multi},w}(Y_1, Z_1) | w_1, \ldots, w_d], \\ & \mathbb{E}_P[g^2_{\text{Multi},w}(Z_1, Z_2) | w_1, \ldots, w_d] \big\}.\end{aligned}$$

Moreover each conditional expected value is computed as

$$\begin{aligned}
\mathbb{E}_P[g^2_{\mathrm{Multi},w}(Y_1,Y_2)|w_1,\dots,w_d] &= \sum_{k=1}^{d} w_k^{-2} p_Y^2(k), \\
\mathbb{E}_P[g^2_{\mathrm{Multi},w}(Z_1,Z_2)|w_1,\dots,w_d] &= \sum_{k=1}^{d} w_k^{-2} p_Z^2(k), \\
\mathbb{E}_P[g^2_{\mathrm{Multi},w}(Y_1,Z_1)|w_1,\dots,w_d] &= \sum_{k=1}^{d} w_k^{-2} p_Y(k) p_Z(k) \\
&\le \frac{1}{2}\sum_{k=1}^{d} w_k^{-2} p_Y^2(k) + \frac{1}{2}\sum_{k=1}^{d} w_k^{-2} p_Z^2(k) \\
&\le \max\left\{ \sum_{k=1}^{d} w_k^{-2} p_Y^2(k),\ \sum_{k=1}^{d} w_k^{-2} p_Z^2(k) \right\}.
\end{aligned}$$

This leads to

$$\psi_{YZ,2}(P) \le \max\left\{ \sum_{k=1}^{d} w_k^{-2} p_Y^2(k),\ \sum_{k=1}^{d} w_k^{-2} p_Z^2(k) \right\}.$$

G.21 Proof of Proposition 8.11

We note that the test statistic considered in Proposition 8.11 is essentially the same as that considered in Proposition 8.10 with different weights. Hence following the same line of the proof of Proposition 8.10, we may arrive at the point (G.47) where the type II error of the considered permutation test is at most β when

$$\epsilon_n^2 \ge \frac{C}{\beta}\log\left(\frac{1}{\alpha}\right)\max\left\{ \sqrt{\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2}\frac{p_Y^2(k)p_Z^2(k)}{w_{k_1,k_2}^2}}\frac{1}{n},\ \frac{d_1^{1/2}d_2^{1/2}}{n} \right\}. \tag{G.50}$$

Similarly as before, let us remove the randomness from $w_{1,1},\dots,w_{d_1,d_2}$ by applying Markov's inequality. First recall that $m_1 = \min\{n/2, d_1\}$ and $m_2 = \min\{n/2, d_2\}$ and thus

$$\begin{aligned}
w_{k_1,k_2} &= \left[\frac{1}{2d_1} + \frac{1}{2m_1}\sum_{i=1}^{m_1}\mathbb{1}(Y_{3n/2+i} = k_1)\right] \times \left[\frac{1}{2d_2} + \frac{1}{2m_2}\sum_{j=1}^{m_2}\mathbb{1}(Z_{5n/2+i} = k_2)\right] \\
&\le \frac{1}{4d_1d_2}\left[1 + \sum_{i=1}^{m_1}\mathbb{1}(Y_{3n/2+i} = k_1)\right] \times \left[1 + \sum_{j=1}^{m_2}\mathbb{1}(Z_{5n/2+i} = k_2)\right].
\end{aligned}$$

Based on this observation, we have

$$\begin{aligned}
&\mathbb{E}_P\left[\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2}\frac{p_Y^2(k)p_Z^2(k)}{w_{k_1,k_2}^2}\right] \\
\leq\ & 16d_1^2d_2^2\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2}\mathbb{E}_P\left[\frac{p_Y^2(k)p_Z^2(k)}{\{1+\sum_{i=1}^{m_1}\mathbb{1}(Y_{3n/2+i}=k_1)\}^2\{1+\sum_{j=1}^{m_2}\mathbb{1}(Z_{5n/2+i}=k_2)\}^2}\right] \\
\leq\ & 16d_1^2d_2^2\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2}\mathbb{E}_P\left[\frac{p_Y^2(k)p_Z^2(k)}{\{1+\sum_{i=1}^{m_1}\mathbb{1}(Y_{3n/2+i}=k_1)\}\{1+\sum_{j=1}^{m_2}\mathbb{1}(Z_{5n/2+i}=k_2)\}}\right] \\
\overset{(i)}{\leq}\ & 16d_1^2d_2^2\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2}\frac{p_Y^2(k)p_Z^2(k)}{(m_1+1)(m_2+1)p_Y(k)p_Z(k)} \\
\leq\ & \frac{16d_1^2d_2^2}{(m_1+1)(m_2+1)},
\end{aligned}$$

where (i) uses the independence between $\{Y_{3n/2+1},\ldots,Y_{3n/2+m_1}\}$ and $\{Z_{5n/2+1},\ldots,Z_{5n/2+m_2}\}$ and also the inverse binomial moment in (G.48). Therefore Markov's inequality yields

$$\begin{aligned}
\mathbb{P}_P\left(\sqrt{\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2}\frac{p_Y^2(k)p_Z^2(k)}{w_{k_1,k_2}^2}}\geq t\right) &\leq \frac{1}{t^2}\mathbb{E}_P\left[\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2}\frac{p_Y^2(k)p_Z^2(k)}{w_{k_1,k_2}^2}\right] \\
&\leq \frac{16d_1^2d_2^2}{t^2(m_1+1)(m_2+1)}.
\end{aligned}$$

This implies that with probability at least $1-\beta$, we have

$$\sqrt{\sum_{k_1=1}^{d_1}\sum_{k_2=1}^{d_2}\frac{p_Y^2(k)p_Z^2(k)}{w_{k_1,k_2}^2}}\leq\frac{4d_1d_2}{\sqrt{\beta(m_1+1)(m_2+1)}}.$$

Under this event, condition (G.50) is implied by

$$\epsilon_n^2\geq\frac{C_1}{\beta^{3/2}}\log\left(\frac{1}{\alpha}\right)\max\left\{\frac{d_1d_2}{m_1^{1/2}m_2^{1/2}n},\ \frac{d_1^{1/2}d_2^{1/2}}{n}\right\}.$$

By putting the definition of $m_1=\min\{n/2,d_1\}$ and $m_2=\min\{n/2,d_2\}$ where $d_1\leq d_2$ and noting that $\epsilon_n^2\leq 4$, the condition is further implied by

$$\epsilon_n^2\geq\frac{C_2}{\beta^{3/2}}\log\left(\frac{1}{\alpha}\right)\max\left\{\frac{d_1^{1/4}d_2^{1/2}}{n^{3/4}},\ \frac{d_1^{1/2}d_2^{1/2}}{n}\right\},$$

for a sufficiently large $C_2 > 0$. The remaining steps are exactly the same as those in the proof of Proposition 8.10. This completes the proof of Proposition 8.11.

G.22 Proof of Proposition 8.12

The proof of Proposition 8.12 is motivated by Meynaoui et al. (2019) who study the uniform separation rate for the HSIC test. In contrast to Meynaoui et al. (2019) who use the critical value based on the (unknown) null distribution, we study the permutation test base on the MMD statistic. The structure of the proof is as follows. We first upper bound $\psi_{Y,1}(P)$, $\psi_{Z,1}(P)$ and $\psi_{YZ,2}(P)$ to verify the sufficient condition given in Lemma G.0.1. We then provide a connection between the expected value of the MMD statistic and L_2 distance $\|f_Y - f_Z\|_{L_2}$. Finally, we conclude the proof based on the previous results. Throughout the proof, we write the Gaussian kernel $K_{\lambda_1,\ldots,\lambda_d,d}(x-y)$ in (8.44) as $K_{\lambda,d}(x-y)$ so as to simplify the notation.

• **Verification of condition (G.8).** In this part of the proof, we find upper bounds for $\psi_{Y,1}(P)$, $\psi_{Z,1}(P)$ and $\psi_{YZ,2}(P)$. Let us start with $\psi_{Y,1}(P)$. Recall that $\psi_{Y,1}(P)$ is given as

$$\psi_{Y,1}(P) = \mathrm{Var}_P\{\mathbb{E}_P[\overline{h}_{\mathrm{ts}}(Y_1, Y_2; Z_1, Z_2)|Y_1]\},$$

where $\overline{h}_{\mathrm{ts}}$ is the symmetrized kernel (8.7). Using the definition, it is straightforward to see that

$$\begin{aligned}\psi_{Y,1}(P) = &\ \mathrm{Var}_P\{\mathbb{E}_P[g_{\mathrm{Gau}}(Y_1, Y_2)|Y_1] - \mathbb{E}_P[g_{\mathrm{Gau}}(Y_1, Z_1)|Y_1]\} \\ \leq &\ \mathbb{E}_P[\{\mathbb{E}_P[g_{\mathrm{Gau}}(Y_1, Y_2)|Y_1] - \mathbb{E}_P[g_{\mathrm{Gau}}(Y_1, Z_1)|Y_1]\}^2].\end{aligned}$$

Let us denote the convolution $f_Y - f_Z$ and $K_{\lambda,d}$ by

$$(f_Y - f_Z) * K_{\lambda,d}(x) = \int_{\mathbb{R}^d} [f_Y(t) - f_Z(t)] K_{\lambda,d}(x-t)dt,$$

where $K_{\lambda,d}$ can be recalled from (8.44). Then the upper bound of $\psi_{Y,1}(P)$ is further bounded by

$$\begin{aligned}\mathbb{E}_P[\{\mathbb{E}_P[g_{\mathrm{Gau}}(X_1, X_2)|X_1] - \mathbb{E}_P[g_{\mathrm{Gau}}(X_1, Y_1)|X_1]\}^2] = &\ \int_{\mathbb{R}^d} f_Y(x)\big[(f_Y - f_Z) * K_{\lambda,d}(x)\big]^2 dx \\ \leq &\ \|f_Y\|_\infty \|(f_Y - f_Z) * K_{\lambda,d}\|_{L_2}^2.\end{aligned}$$

By symmetry, $\psi_{Z,1}(P)$ can be similarly bounded. Thus

$$\begin{aligned}\psi_{Y,1}(P) &\leq \|f_Y\|_\infty \|(f_Y - f_Z) * K_{\lambda,d}\|_{L_2}^2, \\ \psi_{Z,1}(P) &\leq \|f_Z\|_\infty \|(f_Y - f_Z) * K_{\lambda,d}\|_{L_2}^2.\end{aligned} \tag{G.51}$$

Moving onto $\psi_{YZ,2}(P)$, we need to compute $\mathbb{E}_P[g_{\text{Gau}}^2(Y_1, Y_2)]$, $\mathbb{E}_P[g_{\text{Gau}}^2(Z_1, Z_2)]$ and $\mathbb{E}_P[g_{\text{Gau}}^2(Y_1, Z_1)]$. Note that

$$K_{\lambda,d}^2(x) = \frac{1}{(4\pi)^{d/2}\lambda_1 \cdots \lambda_d} K_{\lambda/\sqrt{2},d}(x),$$

where $K_{\lambda/\sqrt{2},d}(x)$ is the Gaussian density function (8.44) with scale parameters $\lambda_1/\sqrt{2}, \ldots, \lambda_d/\sqrt{2}$. Therefore it can be seen that

$$\begin{aligned}\mathbb{E}_P[g_{\text{Gau}}^2(Y_1, Y_2)] &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K_{\lambda,d}^2(y_1 - y_2) f_Y(y_1) f_Y(y_2) dy_1 dy_2 \\ &= \frac{1}{(4\pi)^{d/2}\lambda_1 \cdots \lambda_d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K_{\lambda/\sqrt{2},d}(y_1 - y_2) f_Y(y_1) f_Y(y_2) dy_1 dy_2 \\ &\leq \frac{\|f_Y\|_\infty}{(4\pi)^{d/2}\lambda_1 \cdots \lambda_d}\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d} K_{\lambda/\sqrt{2},d}(y_1 - y_2) dy_1\right] f_Y(y_2) dy_2 \\ &\leq \frac{M_{f,d}}{(4\pi)^{d/2}\lambda_1 \cdots \lambda_d},\end{aligned}$$

where $\max\{\|f_Y\|_\infty, \|f_Z\|_\infty\} \leq M_{f,d}$. The other two terms $\mathbb{E}_P[g_{\text{Gau}}^2(Z_1, Z_2)]$ and $\mathbb{E}_P[g_{\text{Gau}}^2(Y_1, Z_1)]$ are similarly bounded. Thus we have

$$\psi_{YZ,2}(P) \leq \frac{M_{f,d}}{(4\pi)^{d/2}\lambda_1 \cdots \lambda_d}. \tag{G.52}$$

Given bounds (G.51) and (G.52), Lemma G.0.1 shows that the type II error of the considered permutation test is at most β when

$$\begin{aligned}\mathbb{E}_P[U_{n_1,n_2}] \geq\ & C_1(M_{f,d}, d)\sqrt{\frac{\|(f_Y - f_Z) * K_{\lambda,d}\|_{L_2}^2}{\beta}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ &+ \frac{C_2(M_{f,d}, d)}{\sqrt{\lambda_1 \cdots \lambda_d}}\frac{1}{\sqrt{\beta}}\log\left(\frac{1}{\alpha}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).\end{aligned} \tag{G.53}$$

• **Relating $\mathbb{E}_P[U_{n_1,n_2}]$ to L_2 distance.** Next we related the expected value of U_{n_1,n_2} to L_2 distance between f_Y and f_Z. Based on the unbiasedness property of a U-statistic, one can easily verify that

$$\begin{aligned}
\mathbb{E}_P[U_{n_1,n_2}] &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K_{\lambda,d}(t_1-t_2)[f_Y(t_1)-f_Z(t_1)][f_Y(t_2)-f_Z(t_2)]dt_1dt_2 \\
&= \int_{\mathbb{R}^d}[f_Y(t_2)-f_Z(t_2)](f_Y-f_Z)*K_{\lambda,d}(t_2)dt_2 \\
&= \frac{1}{2}\|f_Y-f_Z\|_{L_2}^2+\frac{1}{2}\|(f_Y-f_Z)*K_{\lambda,d}\|_{L_2}^2 \\
&\quad -\frac{1}{2}\|(f_Y-f_Z)-(f_Y-f_Z)*K_{\lambda,d}\|_{L_2}^2.
\end{aligned} \tag{G.54}$$

where the last equality uses the fact that $2xy = x^2+y^2-(x-y)^2$.

• **Completion of the proof.** We now combine the previous results (G.53) and (G.54) to conclude the result. To be more specific, based on equality (G.54), it is seen that condition (G.53) is equivalent to

$$\begin{aligned}
\|f_Y-f_Z\|_{L_2}^2 \;\geq\; & \|(f_Y-f_Z)-(f_Y-f_Z)*K_{\lambda,d}\|_{L_2}^2 \\
& -\|(f_Y-f_Z)*K_{\lambda,d}\|_{L_2}^2 \\
& +C_3(M_{f,d},d)\sqrt{\frac{\|(f_Y-f_Z)*K_{\lambda,d}\|_{L_2}^2}{\beta}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \\
& +\frac{C_4(M_{f,d},d)}{\sqrt{\lambda_1\cdots\lambda_d}}\frac{1}{\sqrt{\beta}}\log\left(\frac{1}{\alpha}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right).
\end{aligned} \tag{G.55}$$

Based on the basic inequality $\sqrt{xy} \leq x+y$ for $x,y \geq 0$, we can upper bound the third line of the above equation as

$$\begin{aligned}
C_3(M_{f,d},d)\sqrt{\frac{\|(f_Y-f_Z)*K_{\lambda,d}\|_{L_2}^2}{\beta}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \;\leq\; & \frac{C_5(M_{f,d},d)}{\beta}\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \\
& +\|(f_Y-f_Z)*K_{\lambda,d}\|_{L_2}^2.
\end{aligned}$$

Therefore the previous inequality (G.55) is implied by

$$\begin{aligned}
\epsilon_{n_1,n_2}^2 \;\geq\; & \|(f_Y-f_Z)-(f_Y-f_Z)*K_{\lambda,d}\|_{L_2}^2 \\
& +\frac{C(M_{f,d},d)}{\beta\sqrt{\lambda_1\cdots\lambda_d}}\log\left(\frac{1}{\alpha}\right)\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right),
\end{aligned}$$

where we used the condition $\prod_{i=1}^d \lambda_i \leq 1$. This completes the proof of Proposition 8.12.