Non-Parametric Causal Discovery for Discrete and Continuous Data

Submitted in partial fulfillment of the

requirements of the degree of

Doctor of Philosophy

in

Statistics & Data Science

and

Engineering & Public Policy

Octavio César Mesner

B.S., Mathematics and Philosophy, John Carroll University
M.S., Mathematics, John Carroll University
M.S., Biostatistics, Georgetown University
M.S., Machine Learning, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA
August 2020

Acknowledgements

This journey has been a challenge and would not have been possible without the support of many people who have given me their time and encouragement. Cosma Shalizi, my primary advisor, gave me the freedom to pursue my own interests but direction when I was stuck. Larry Wasserman's insight and breath of knowledge steered many technical aspects of this research. Kun Zhang gave me formal training in causal discovery methods. Alex Davis pushed me to strive for my best and think more critically on the intersection of data method and policy analysis. This dissertation would be not possible without Liz Casman. Even before I was admitted, Liz took the time understand my background and potential, even though my application was different from most of the students admitted to PhD programs at CMU. Because of this, Liz has single-handedly changed my life and I am infinitely grateful.

The StatDS and EPP communities at CMU have cultivated so many relationships for me. The EPP PhD students inspire me with their advocacy through research. The StatDS PhD students amaze me with their creativity for data methods and analysis. The faculty in both departments provide an example of what we can aspire to with strategic collaboration and hard work. The EPP and StatDS staff have provided so much support over the years, helping me to navigate new and unknown administrative issues and just making CMU a warmer place than it would be otherwise.

CMU's Center for Student Diversity and Inclusion cultivated a community of smart, tireless advocates of peoples who are frequently marginalized in US society. CMQ+ has worked consistently to unite and advocate for CMU's graduate LGBTQ+ communities. BGSO and LGSA have provided spaces for the Black and Latinx graduate students to remind the University that it shares in the responsibility of understanding, advocating for, and promoting individuals from our communities.

I am proud to be associated with these groups.

This PhD would not have been possible without funding. The College of Engineering Dean's Fellowship funded the first year of my PhD. Tamar Krishnamurti, through a successful grant to the Center for Machine Learning and Health, funded my second and third years. The Department of Statistics and Data Science funded my fourth year. My last year was funded through the Ford Foundation Dissertation Fellowship. Thank you.

Before CMU, Borromeo Seminary, John Carroll University, Georgetown University, and the Infectious Disease Clinical Research Program all prepared me for the interdisciplinary research that I always aspired to do. Brother Charlie McElroy taught me to reach beyond what I thought was possible, helping me write my first scholarship applications to the Hispanic Scholarship Fund and the Congressional Hispanic Caucus Institute. With his guidance, I received both. Doctors Barbara D'Ambrosia, Paul Shick, and Patrick Chen taught me to construct mathematical proofs. Doctor Francoise Seillier-Moiseiwitsh showed me how to be a conscientious biostatistician.

Finally, my family and friends have been immensely encouraging, caring, and understanding during and before my PhD. My mother immigrated to the United States from Nicaragua while she was pregnant with me during the Nicaraguan Revolution in the 1980s. Her sacrifices have given me so many opportunities. My friends Gustavo Alverio, Freddy Herrera, Greg Taylor, Jay Stetz, and Rahul Ladhania have stuck with me through countless predicaments. Our friendships have given me so much joy. My siblings, Krysten, Jack, Octavio José, Marta, and Blanca have all taken routes in their lives that make me proud to be their brother. My husband, Antoine, with his combination of work ethic, kindness, passion, and intelligence, inspires me to be the best human being I can.

Abstract

Subject-matter experts typically think of their datasets as causes and effects between many variables, forming a large, complex causal system. Directed acyclic graphs (DAG), also called Bayesian networks, provide a natural way to conceptualize these systems. In contrast, regression modeling can provide strong evidence for the local, causal neighborhood of an outcome within the causal system, but providing structure for the larger system is challenging with regression. Despite its value as exploratory data analysis or in conjunction with regression models to refine causal understanding, methods for estimating the causal structure underlying a dataset, *causal discovery*, are rare in fields such as epidemiology, possibly due to the difficulty handling data with continuous and discrete random variables.

This thesis focuses on developing a causal discovery method for researchers whose data typically are comprised of both discrete and continuous variables. Its primary contribution is the development of an estimator for *graph divergence*, the Kullback-Leibler divergence between the full, joint distribution and the Bayesian factorization indicated by a DAG. Graph divergence is a generalization of conditional mutual information: it quantifies the fit of a DAG to the data, with greater divergence indicating worse fit and a divergence of zero indicating a perfect characterization of the conditional independence relationships among the variables. Its nearest neighbor approach gives the estimator the capability to handle mixed data. We show that the estimator is consistent and its convergence separately for the continuous and discrete case under some assumptions.

Last, we demonstrate a way to use graph divergence with a greedy Markov equivalence search algorithm in practice. Though this work is not complete, we estimate causal relationships between personal demographics, sexual risk behaviors, and HIV Pre-exposure prophylaxis among men who have sex with men (MSM) on the American Men's Internet Survey data. This work may be able to inform public health initiatives and guidelines surrounding sexual health of MSM.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

As I write this introduction, the world is in the midst of a global pandemic. Scientists have scrambled to study many different aspects of the virus, SARS-CoV-2, and its impact on the economy, education, personal well-being, global relations, the climate, and so on. Of particular interest to policymakers and citizens alike is learning what can be done to limit transmission risk. Even knowing where to start looking can be hard. One characteristic that makes infectious disease challenging is that modes of transmission vary by circumstance, population, or even evolve over time. Discerning combinations of circumstances that will lead to transmission can be difficult partly because patterns of sequential events do not always indicate causation and partly because its seemingly probabilistic nature. This problem is compounded with increasing numbers of potential risk factors at play. During the emergence of SARS-CoV-2, no one possessed specific prior knowledge on its prolific spread. Causal discovery, data methods that do not require prior causal understanding to sort through, evaluate, and organize risk factors into an causal diagram, have potential to supplement expert opinion. This could be particularly helpful in complex, and causally unstructured problems, such as emerging infectious disease. This thesis focuses on causal discovery for the types of data found in epidemiological settings, where data are frequently comprised of a mix between discrete and continuous variables whose statistical dependence may or may not have a linear relationship.

Working as a biostatistician in HIV and sexually transmitted infections (STI) research, my primary role was to understand a causal theory then investigate it with statistical models, primarily with regression. In my experience, medical doctors and epidemiologists would explain their theories as chains of events, couched within a mental model more complex than I could capture with regression. But, understanding the larger picture was necessary for variable selection, which in practice tended to be more of trial and error. This experience and the desire to understand the data in a more principled and systematic way caused me to look for data-driven approaches for variable selection and causal discovery. During this time, I also be-

came familiar with some limitations with statistical tools resulting when data are collected from people such as missing or censored data, mixed variables types, non-linearity, followup variation, and so on. These experiences are what inspired this line of research.

This thesis shows my journey of attempts to contribute to this goal and the understanding that I developed throughout the process. Chapter 2 gives a brief background on prior methods in causal discovery. For the following work both constraint-based methods such as the PC algorithm and score-based methods like GES were my primary starting point.

Chapter 3 is a study using a causal discovery method to illustrate causal pathways from maternal stress during pregnancy to pregnancy outcomes. This topic was interesting because the maternal stress during pregnancy is widely believed to contribute to preterm birth but the causal pathways are not well understood. For this study, we wrote an implementation of kernel conditional independence test in R. With little experience in software development, writing this code took a significant amount of time. We used this code with the PC algorithm from the pcalg R package to do the data analysis. As my first time doing research with casual discovery, I learned a lot on how researchers experience and understand graphical models which informed my work moving forward.

Chapter 4 presents a method for calculating conditional mutual information with discrete and continuous data. Information theory in general seemed to be a promising path forward because there were already estimators for discrete data and continuous data, but not both. Moreover, information estimators quantify dependence/independence generally, as opposed to linear correlation. In this paper, we show consistency but do not give a convergence rate. Under the assumptions we made, it is possible to have arbitrarily slow convergence. Though, the simulations we used performed very well.

Chapter 5 expands on Chapter 4, presenting a method to estimate the Kullback-

Leibler divergence between a Bayesian factorization of a joint distribution and the true distribution, called graph divergence. This work is unfinished and will likely be split into two or three papers. From what is already finished, it seems possible that we can prove that the continuous estimator obeys the central limit theorem (CLT) with specific assumptions. This result would deserve to be its own paper. Following this, I would like to continue work on a mixed graph divergence estimator with assumptions to support another CLT theorem. It is clear from the CMI work that the mixed graph divergence estimator is similarly consistent in this work. Moving to a complete causal discovery algorithm, it will be necessary to understand how this estimator performs with a greedy DAG search algorithm given that is has no sparsity penalty as currently implemented.

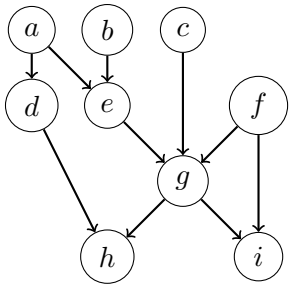
Chapter 6 is preliminary work using the method from Chapter 5 combined with a graph search algorithm to causally organize risk factors contributing sexually transmitted infections among men who have sex with men. For the thesis, I felt it was important to have a preliminary results as to show that this method does (or does not) have potential in practice. But, as stated, I would like to take more time to consider alteration to both the graph divergence estimator and the greedy DAG search algorithm. I would also like to consider ways to communicate estimation error to subject-matter experts who are not statisticians or data scientists.

Chapter 2

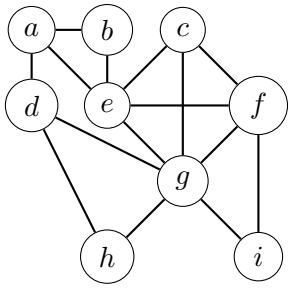
Background

2.1 Graph Learning Algorithms

Graph learning algorithms typically attempt to build either a directed acyclic graph (DAG) or a conditional independence graph (CIG). DAGs, sometimes called Bayesian networks, show all causal relationships between variables including the direction of causation; see figure 2.1a. CIGs show all dependencies conditioning on every other variable in the data without showing the direction of causation; see figure 2.1b. For any DAG, there is a corresponding CIG by connecting nodes that have a common cause, though the converse is not true, figure 2.1 shows this correspondence.



(a) Directed Acyclic Graph



(b) Conditional Independence Graph

Figure 2.1: Types of graphical models

Marginal dependence between variables X and Y ($X \not\perp\!\!\!\perp Y$) is thought to be induced in three ways: causation ($X \to Y$ or $Y \to X$), confounding ($X \leftarrow Z \to Y$), or sampling bias ($X \to S \leftarrow Y$ and S influences sampling). Thus, anytime a marginal dependence between variables is observed, it is thought to be induced by at least one of these. Considering conditional dependence relationships is also helpful. For an undirected path of three variables $X - Z - Y$, there are 4 possible causal structures, see table 2.1.

Moving in the other direction, learning structure from data, it is necessary to assume *faithfulness* for most methods. Faithfulness assumes that if there is a causal relationship in the data, it will exhibit dependence. This may seem straightforward but it is possible to have a causal relationship between variables but still observe independence in the variables. In general, structure learning is computationally

Table 2.1: Causal structures for three variables with marginal and conditional independence

	Name	Structure	Marginal	Conditional
1.	Causal Trail	$X \to Z \to Y$	$X \not\perp\!\!\!\perp Y$	$X \perp\!\!\!\perp Y \mid Z$
2.	Evidential Trail	$X \leftarrow Z \leftarrow Y$	$X \not\perp\!\!\!\perp Y$	$X \perp\!\!\!\perp Y \mid Z$
3.	Common Cause	$X \leftarrow Z \to Y$	$X \not\perp\!\!\!\perp Y$	$X \perp\!\!\!\perp Y \mid Z$
4.	Common Effect	$X \to Z \leftarrow Y$	$X \perp\!\!\!\perp Y$	$X \not\perp\!\!\!\perp Y \mid Z$

challenging because the discrete search space, the set of all graphs on p variables, grows super-exponentially with p [1]. The remaining part of this section explains various methods for structure learning.

2.2 Constraint-Based Method/PC

Using the principals above, the PC algorithm [2] uses a series of tests to determine conditional independence relationships between variables in order to build a partial DAG where some edges may not be oriented. Intuitively, PC begins by forming a complete graph, connecting each node to every other node. Using an independence test (PC does not specify one), PC removes any edges whose nodes/variables test not significant marginally. In the following steps, it exhaustively checks conditional independence relationships for variables that remain connected by and edge, conditioning on neighbors of either. Algorithm 1 modified from reference [3] shows how PC builds the DAG skeleton. The following step in PC shows how to orient edges

based on detecting common effect conditional independence patterns in the graph.

Algorithm 1: PC Skeleton

Data: Vertex Set V, Conditional Independence Information

Result: Estimated skeleton graph

Form complete undirected graph $\tilde{C}$ on the vertex set V;

Set $\ell = 0$ (size of conditioning set) and $C = \tilde{C}$ (current graph);

while *for each ordered pair of adjacent nodes* $i, j : |adj(C, i)\backslash\{j\}| < \ell$ **do**

  Choose a (new) ordered pair of nodes i, j that are adjacent in C and that $\text{adj}(C, i)\backslash\{j\} \geq \ell$;

  for *all ordered pairs of adjacent variables* i *and* j *such that* $i, j : |adj(C, i)\backslash\{j\}| \geq \ell$ *and* $K \subseteq adj(C, i)\backslash\{j\}$ *with* **do**

    while *edge* i, j *exists or there is an untested* $K \subseteq adj(C, i)\backslash\{j\}$ *with* $|K| = \ell$ **do**

      Choose a set of variables K such that $K \subseteq \text{adj}(C, i)\backslash\{j\}$ with $|K| = \ell$ and test $i \perp\!\!\!\perp j|K$;

      if $i \perp\!\!\!\perp j|K$ **then**

        Delete edge i, j;

      end

    end

  end

end

2.3 Score-Based Methods/GES

Score-based methods are a class of graph-learning algorithms that, rather than directly using conditional independence tests, use a score function to measure a graph's fit to the data. The idea is that graphs that fit the data poorly will not capture the correct independence relationships in the data. In this section, we mainly consider greedy equivalence search (GES) from reference [4].

Score-based methods typically build on parametric regression model selection

techniques based on scoring such as AIC or BIC. Recall that the BIC of a model is defined as BIC $= \ell(\widehat{\theta}) - \frac{d}{2}\log n$ where $\ell(\widehat{\theta})$ is the log likelihood, d is the number of parameters, and n is the number of observations. The model with the greatest score is thought to be the best model among the models considered because it attempts to maximize log-likelihood while accounting for using the data twice (once to estimate the parameters and once to estimate the log-likelihood). More recently, a reproducing kernel Hilbert space (RKHS) estimator has been developed which relaxes many of the assumptions for AIC and BIC [5].

DAG estimation relies on the fact that DAGs encode a joint distribution factorization, chapter 3 from reference [6]. That is, given the causal DAG, $\mathcal{G}$, a factorization of its corresponding joint distribution, $P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i|\text{pa}_{\mathcal{G}}(X_i))$ where $\text{pa}_{\mathcal{G}}(X_i)$ is the set of parent nodes of X_i in $\mathcal{G}$. Thus, the joint log-likelihood can be decomposed into the sum of log-likelihoods of regression models for each $P(X_i|\text{pa}_{\mathcal{G}}(X_i))$. However, rather than adding log-likelihoods to generate the composite score for any given graph, we add BICs together. This has the advantage of more accurately modeling DAG fit than likelihood alone.

Unfortunately, as mentioned in section 2.2, finding the graph that maximizes this composite score involves searching the space of DAGs on p nodes is challenging if p is large. Further, because distinct DAGs can have identical conditional independence relationships between variables as described in reference [4] and chapter 3 from reference [6]. DAGs with identical conditional independence relationships form an equivalence class whose composite scores are all the same as well. To simplify the search process, GES defines graphs *adjacent* to $\mathcal{G}$ as those graphs that can be obtained by adding an edge, removing an edge, or changing an edge orientation. Only considering equivalence classes of DAG, the search space is narrowed, simplifying the search. In a greedy search, from any DAG equivalence class, we score all adjacent equivalence classes and choose the class the gives the largest score. Reference [4] suggests we use a two phases process. In the first phase, we start with an

empty DAG, adding edges until a local maximum is reached. In the second phase, we removed edges again until a local maximum is reached, at with point we output the result.

2.4 LiNGAM

Unlike constraint-based and score-based DAG learning techniques, Linear non-Gaussian acyclic models [7] eponymously assumes that variables are continuous but do not have a Gaussian distribution and that the generating process for variable x_i is a linear function of its parents, $x_i = \sum_{x_j \in \text{pa}(x_i)} b_{ij} x_j + \epsilon_i$ and ϵ_i is residual noise such that $x_i \perp\!\!\!\perp \epsilon_i$ and ϵ_i is *non-Gaussian.*

Using this paradigm, we can write $X = BX + \epsilon$ where X is the vector of variables, $x_1, \ldots, x_d$, and ϵ is the vector containing the entries $\epsilon_1, \ldots, \epsilon_d$ and all independently generated. Notice that because we are assuming that the underlying structure is a DAG, the columns and rows of B can be simultaneously permuted to create a strictly upper or lower triangular matrix. Otherwise, there would be feedback loops in the DAG. Solving for X, we have $X = (I - B)^{-1}\epsilon$. Using independent component analysis (ICA) [8], we can search for a matrix $A = (I - B)^{-1}$ that attempts to find linear transformations of the entries of ϵ which are as independent as possible. Note, however, that A is unique up to scaling and permutation as discussed above [9]. The requirement of non-Gaussianity relies on ICA's requirement of non-Gaussianity.

For LiNGAM, the ordering of rows and columns of B, and thus, A is important as well as the scaling. Once ICA is applied to the data, the LiNGAM algorithm works exclusively with $W = A^{-1}$. The goal of the LiNGAM algorithm is to transform W to an lower triangular matrix with ones along the diagonal. Reference [7] summarizes the algorithms as follows:

1. Given an $m \times n$ data matrix X with $(m \ll n)$, where each column contains one sample vector x, first subtract the mean from each row of X, then apply ICA to

obtain a factorization $X = A\epsilon$ where ϵ contains the independent components of X.

2. Find the one and only permutation of rows of W which yields a matrix $\tilde{W}$ without any zeros on the main diagonal. In practice, small estimation errors will cause all elements of W to be non-zero, and hence we search for the permutation that minimizes $\sum_i \frac{1}{|\tilde{W}_{ii}|}$.

3. Divide each row of $\tilde{W}$ by its corresponding diagonal element, to yield a new matrix $\tilde{W}'$ with all ones on the diagonal.

4. Compute an estimate $\widehat{B}$ of B using $\widehat{B} = I - W'$.

5. To find a causal ordering, find the permutations matrix P applied equally to rows and columns of $\widehat{B}$ which yields matrix $\tilde{B} = P\widehat{B}P^T$ which is as close to possible to the strictly lower triangular measured by $\sum_{i \leq j} \tilde{B}_{ij}^2$.

2.5 Glasso

Graphical lasso or Glasso [10] attempts to estimate the inverse covariance matrix, Σ^{-1}, of a multivariate Gaussian distribution. Reference [11] shows that a zero entry in the ijth entry of Σ^{-1} implies that the i and j components of the random vector are conditionally independent given all other components. That is, in a CIG or Markov network, nodes i and j will be connected iff the ijth entry of Σ^{-1} is non-zero. Unfortunately, it is unlikely for most estimates, $\widehat{\Sigma}^{-1}$, of Σ^{-1} to contain components that are exactly zero. Reference [12] used an ℓ_1 penalty on the log-likelihood estimate for the inverse covariance matrix to enforce zero matrix entries, changing the problem to a semidefinite programming problem:

$$\widehat{\Sigma}^{-1} = \underset{\Theta \succ 0}{\arg\max} \left[\log\det\Theta - \operatorname{tr}\left(\widehat{\Sigma}\Theta\right) - \lambda \left\|\Theta\right\|_1 \right]$$

where $\widehat{\Sigma} = \frac{1}{N}\sum_i (x_i - \widehat{\mu})(x_i - \widehat{\mu})^T$, λ is the regularization parameter to be tuned, and $\|\cdot\|_1$ is the ℓ_1 norm, the sum of absolute values of the matrix entries.

This model is problematic for most datasets because it is not likely that each variable is Gaussian, or even continuous. Further, this model will need to either drop or impute missing values.

2.6 Lasso Neighborhood Selection

Lasso neighborhood selection [13] aims to build a CIG by learning the neighborhood (set of adjacent nodes) for each node/variable in the graph. The framework assumes a p-dimensional multivariate Gaussian distribution where neither the number of observations nor the number of variables is fixed. Using the Markov property of CIGs, we have that any variable not in the Markov blanket of a target variable will be conditionally independent of the target given the Markov blanket. Said differently, for any node, a, in the graph, and for its corresponding variable X_a, consider the vector θ^a such that

$$\theta^a = \underset{\theta:\theta_a=0}{\arg\min}\, \mathbb{E}\left(X_a - \sum_{1\leq i\leq p} \theta_i X_i\right)^2.$$

We would expect $\theta_i = 0$ only for X_i in Markov blanket of X_a. Unfortunately, the parameter estimates for variables not in the Markov blanket of X_a are not likely to be exactly zero for this optimization problem without an additional constraint. Using the ℓ_1 penalty on the previous objective function, given as

$$\widehat{\theta}^{a,\lambda} = \underset{\theta:\theta_a=0}{\arg\min}\left(\|X_a - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_1\right),$$

we can coerce some variables to be exactly zero, yielding a neighborhood estimate of $\left\{k : \widehat{\theta}_k^{a,\lambda} \neq 0\right\}$. Under some assumptions, reference [13], shows that neighborhoods can be estimated. If a neighborhood is estimated for each variable, there may be

conflicts on the existence of an edge that can either be resolved by accepting an edge only when each variable considered the other to be its neighbor OR when one or the other does. A benefit to this approach is it is less computationally intensive than the previous methods and because it is feasible on data where $p \gg n$.

2.7 GRaFo

Graphical random forest (GRaFo) [14] builds a CIG using random forests with stability selection [15]. Random forest variable importance is used here as a measure of association between variables that can handle both continuous and discrete variables well. Briefly, the variable importance of a variable X_i for an outcome Y is calculated for the forest by averaging the difference in prediction accuracy with the out-of-bag sample for each tree, once without permuting X_i and once with permuting X_i. By regressing a target variable on all other variables, we can then rank variables based on variable importance with respect to the target. To create an edge ranking, generate a variable importance ranking with respect to each variable in the data, and for each pair of variables, keep the minimum of the two scores. Using the edge ranking alone, it is not clear where to draw a cutoff for choosing the top q edges.

Stability selection gives a method for choosing q and bounding the expected number of false positives, $\mathbb{E}[V]$, from graphs generated on subsamples of the full data. Edges in the final graph are those that exist in a proportion, π_{thr}, of graphs generated on the subsamples. The idea is that these edges are "sufficiently stable" among the subsamples. Theorem one from [15] relates $\mathbb{E}[V], \pi_{\text{thr}}$ and q under the fairly week assumptions:

$$\mathbb{E}[V] \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p(p-1)/2}$$

where $\pi_{\text{thr}} \in (\frac{1}{2}, 1)$ and p is the number of variables in the data. By specifying $\mathbb{E}[V]$ and π_{thr}, we can choose $q = \lfloor \sqrt{(2\pi_{\text{thr}} - 1)p(p-1)/2} \rfloor$. GraFo works by first gen-

erating n_{sub} subsamples without replacement, $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n_{\text{sub}})}$, each of size $\lfloor n/2 \rfloor$. For each subsample, rank each edge using the variable importance and choose the top q edges. Include in the final graph only those edges that appear in at least π_{thr} proportion of the graphs.

A draw back of variable importance is that it has been shown to not satisfy the data processing inequality [16]. Fixes for this problem have been proposed but none are computationally feasible.

Chapter 3

Maternal Stress and Pregnancy

3.1 Application in Stress and Pregnancy

Many medical risks are multifactorial, with complex pathways connecting causes to adverse events. The biological and psychosocial pathways leading to adverse pregnancy outcomes, such as preterm birth, are complex and only partially understood. In this section, we show how to use structure learning to model those pathways simultaneously, providing an interpretable high-level view of potential causal mechanisms. We examined adverse pregnancy outcomes because they are a well-studied topic, where researchers have examined a range of potential causes, from psychosocial factors such as stress, childhood neglect, and depression, to biological indicators such as inflammation, hypertension, and diabetes.

Maternal stress has been one critical focus of PTB research. While there are theories on biological pathways linking stress and PTB, the evidence is mixed [17]. Dole and others found that anxiety, negative life events, and perceived racial discrimination were all associated with increased risk of PTB [18]. Copper and others found that stress was associated with both PTB and low birth weight [19]. Kramer and others found that only anxiety, out of a large number of stressors and psychological distress measures, was associated with PTB. They further explored biological pathways by sampling stress biomarkers, but were not able to confirm a causal mechanism [20]. In a review of the epidemiology of PTB, Goldenberg and others list stress as one factor among several which can initiate PTB [21]. Interestingly, some studies have also found no association. Glynn and others were not able to predict PTB from anxiety and perceived stress measured at 18 - 20 and 30 - 32 weeks of gestation [22].

3.2 Data

In the Measures of Maternal Stress (MOMS) Study, 744 women were recruited between June 2013 and May 2015 from four sites, Northwestern University, Univer-

sity of Texas Health Science Center at San Antonio, University of Pittsburgh, and Schuylkill County, Pennsylvania, a rural site led by Childrens Hospital of Philadelphia. All women were at least 18 years of age with a singleton intrauterine pregnancy, less than 21 weeks pregnant at enrollment, English-speaking, and with no known fetal congenital anomalies. Enrolled women were examined twice, once between 12 and 21 weeks of gestation (visit A), and again between 32 and 36 weeks of gestation (visit B). Due to a higher proportion of missing data, including some key outcomes, we did not use variables collected at visit B. In all, there were 744 women at visit A and 639 at visit B; ultimately, 686 post-delivery medical records, such as pregnancy outcomes, were available. Additional details about the original data collection can be found in prior publications [23].

3.3 Method

We used the PC algorithm from section 2.2 with the kernel conditional independence (KCI) test [24] for this analysis. In the remainder of the paper, we refer to the method as PC-KCI. We did not use PC to orient edges.

Kernel conditional independence (KCI) test [24] uses the reproducing kernel Hilbert space (RKHS) structure to determine the independence of two random vectors conditioning a third random vector, building on [25] and [26]. This section provides an summary of how Mercer kernels can be used to measure independence. A more detailed account can be found in appendix 3.6.

At a very high overview, kernels are functions that allow mapping probability measures to spaces of functions. Using the machinery in these spaces, it is possible to determine whether probability measures mapped into the function spaces are independent. Let $\mathcal{X}$ be a measurable set and $\mathcal{P}$ be the set of probability measures on $\mathcal{X}$ with corresponding RHKS, $\mathcal{H}$ and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. A kernel

is *characteristic* if the mapping

$$\mu : \mathcal{P} \to \mathcal{H} \text{ defined by } \mathbb{P} \mapsto \int_{\mathcal{X}} k(x, \cdot) d\mathbb{P}(x)$$

is injective. Intuitively, this ensure that $\mathcal{H}$ is a rich enough function space to represent characteristic functions. This is necessary in order to represent characteristic functions of probability measures in $\mathcal{H}$. Recall that random variables $X_1, X_2, \ldots, X_k$ are independent if and only if $\phi_{X_1, X_2, \ldots, X_k}(t) = \prod_{j=1}^{k} \phi_{X_j}(t_j)$ where $\phi_X(t) = \mathbb{E}\left[\exp\left\{itX\right\}\right]$ is the characteristic function of a random variable, X, [27] exercise 3.9.6.

In this application of kernels, however, we use the *cross covariance operator* which is a mapping from one RKHS to another with the property that it can easily compute the covariance of random variable under transformations (functions) in each RKHS. That is, let X, Y, Z be random variables in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ with corresponding RHKSs $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}, \mathcal{H}_\mathcal{Z}$, and kernels $k_\mathcal{X}, k_\mathcal{Y}, k_\mathcal{Z}$, respectively. Assume

$$\mathbb{E}\left[k_\mathcal{X}(X, X)\right] < \infty, \mathbb{E}\left[k_\mathcal{Y}(Y, Y)\right] < \infty, \mathbb{E}\left[k_\mathcal{Z}(Z, Z)\right] < \infty.$$

This ensures that for each random variable, its RKHS is contained within its L^2 space. The *cross covariance operator* is the mapping $\Sigma_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ that satisfies

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

for all $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$. It is called the *covariance operator*, Σ_{XX}, if $X = Y$. The *conditional cross covariance operator* is defined as

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}.$$

The parallel between covariance operators and covariance matrices is that the operators are a generalization of the matrices. These operators are helpful for this work

because if the kernels are characteristics and the operator is the zero mapping, then the random variables must be independent. The Hilbert-Schmidt operator norm, which maps operators to a real scalars so that only the zero operator corresponds to zero, makes it easier to create a statistical test for independence. In short, the test estimates the distribution of the Hilbert-Schmidt norm of the (conditional) cross covariance estimator under the null hypothesis (independence) to determine significance with empirical data.

3.4 Results

We break the results into two parts: affirming prior research and extending prior research. The most obvious relationships are among the neighbors of infant adverse pregnancy outcomes, where gestational weeks is closely related to preterm premature rupture of membranes (PPROM), length of stay in the neonatal intensive care unit (NICU), and percentile gestational weight at birth. Similar patterns emerge between gestational weeks and most maternal adverse outcomes (maternal gestational diabetes, preeclampsia, and gestational hypertension). The positive relationship between BMI, C-Reactive Protein (CRP) and earlier gestational age at delivery confirms established relationships[28, 29, 30]. Similarly, we find that race, a documented risk factor for preterm birth [31, 32], is connected to pre-gestational diabetes, which itself is connected to gestational weeks. In our findings, Hispanic ethnicity is connected to pre-gestational diabetes which we know from the literature to be a population trend.

Of the extending type, we observed that pathways differ by race, such that African-American participants were more likely to have elevated hair cortisol levels, which, in turn, was associated with pre-eclampsia onset and shorter gestational weeks. We also found that higher scores on the Childhood Trauma Questionnaire are associated with small for gestational age infants at birth and, furthermore, that this is related to social problems and perceived social stress in adulthood. Prior

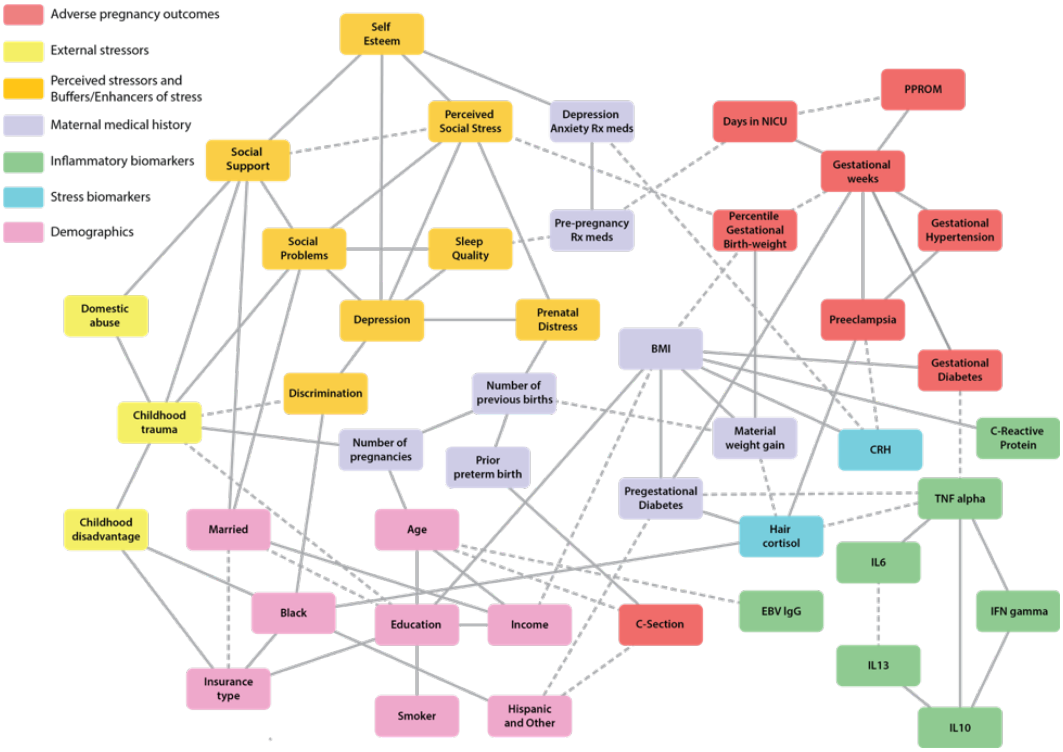


Figure 3.1: Visual representation of results from the algorithm applied to the Measures of Maternal Stress (MOMS) Study. Weeks in which measurement was taken are in parentheses. Solid lines represent edges appearing in both the $p < 0.01$ and $p < 0.05$ graphs. Dashed lines represent edges for $p < 0.05$. The three dotted blue lines connecting TNF alpha to hair cortisol, pregestational diabetes, and gestational diabetes represent key potential false negative edges that may have been missed by PC-KCI

literature has shown a relationship between economically disadvantaged childhood and shorter gestational weeks, when controlling for current income [28]. Here we find that this pathway is related to current economic disadvantage, as indicated by insurance type, suggesting that factors related to economic status that are distinct from income, may be playing a critical role in the preterm birth pathway. Finally, prior preterm birth, age, and Hispanic ethnicity are connected to whether a patient had a C- section, with Hispanic women in the sample almost twice as likely to have a C-section (20%) compared to non-Hispanic women (12%). The graph indicates that

this relationship was not mediated by BMI (a commonly cited cause of C-sections), suggesting other possible explanations, such as patient preference for the procedure or variation between hospitals (e.g., San Antonio vs. Pittsburgh) in C-section rates.

3.5 Discussion

Associations or statistical dependency between two variables are generally induced via one of two ways: direct causation, in which one variable causes the other, or confounding, in which there is a third, unmeasured variable that causes two measured variables. Statistical methods for distinguishing between these cases are limited and often controversial within the statistical community. For this reason, we proceed with caution in making causal claims based solely on observational data. However, the absence of statistical association generally indicates that two variables are not causally linked given that there are enough data to detect an association. The independence tests are probabilistic with an attendant degree of risk of false positives and false negatives. As shown by the failure to detect the pathway between TNF-alpha and gestational weeks, PC-KCI is likely to be too conservative, with a higher chance of false negatives or missing edges between variables in the graph. For analyses based on a priori theoretical models, a regression approach could produce somewhat different results, compared to the exploratory approach of PC-KCI.

3.6 Appendix: KCI

Let $\mathcal{X}$ be a measurable set and $\mathcal{P}$ be the set of probability measures on $\mathcal{X}$ with corresponding RHKS, $\mathcal{H}$ and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. A kernel is *characteristic* if the mapping

$$\mu : \mathcal{P} \to \mathcal{H} \text{ defined by } \mathbb{P} \mapsto \int_{\mathcal{X}} k(x, \cdot) d\mathbb{P}(x)$$

is injective. Intuitively, this ensure that $\mathcal{H}$ is a rich enough function space to represent characteristic functions. This is necessary in order to represent char-

acteristic functions of probability measures in $\mathcal{H}$. Recall that random variables $X_1, X_2, \ldots, X_k$ are independent if and only if $\phi_{X_1,X_2,\ldots,X_k}(t) = \prod_{j=1}^{k} \phi_{X_j}(t_j)$ where $\phi_X(t) = \mathbb{E}\left[\exp\left\{itX\right\}\right]$ is the characteristic function of a random variable, X, [27] exercise 3.9.6. The rest of this section is devoted to the sufficient conditions for ensuring that a kernel is characteristic.

In general, we assume the integrability of all random variables under their respective kernels so that $\mathbb{E}\left[k(X,X)\right] < \infty$. Reference [33] shows that this ensures that $\mu(\mathbb{P}) \in \mathcal{H}$ and that $\mathbb{E}_{\mathbb{P}}\left[f(X)\right] = \langle f, \mu(\mathbb{P})\rangle_{\mathcal{H}}$. Assuming $f \in \mathcal{H}$, we have

$$\begin{aligned} \left|\mathbb{E}_{\mathbb{P}}\left[f(X)\right]\right| &\leq \mathbb{E}_{\mathbb{P}}\left|f(X)\right| && \text{Jensen} \\ &= \mathbb{E}_{\mathbb{P}}\left|\langle f, k(X,\cdot)_{\mathcal{H}}\rangle\right| \\ &\leq \mathbb{E}_{\mathbb{P}}\left[\sqrt{k(X,X)}\left\|f\right\|_{\mathcal{H}}\right] && \text{Cauchy Schwartz} \end{aligned}$$

This shows that $\mathbb{E}_{\mathbb{P}}\left[f(X)\right]$ exists. It's easy to see using Riesz representation theorem that $\mathbb{E}_{\mathbb{P}}\left[f(X)\right] = \langle f, \mu(\mathbb{P})\rangle_{\mathcal{H}}$.

Theorem 3.6.1. *Let $\mathcal{H}$ be an RKHS with corresponding measurable, bounded kernel k on a measurable space $(\mathcal{X}, \mathcal{B})$. If $\mathcal{H} \oplus \mathbb{R}$ (direct sum) is dense in $L^q(\mathcal{X}, \mathbb{P})$ for any $\mathbb{P} \in \mathcal{P}$ and $q \geq 1$, then k is characteristic.*

Proof. To show that μ is injective, assume that $\mu(\mathbb{P}) = \mu(\mathbb{Q})$. Let $\epsilon > 0$ and $A \in \mathcal{B}$. Since $\mathcal{H} \oplus \mathbb{R}$ is dense in $L^q(\mathcal{X}, \mathbb{P})$, there must be a function $f \in \mathcal{H}$ and $c \in \mathbb{R}$ such that $|\mathbb{E}[f(X)] + c - \mathbb{P}(A)| < \epsilon/2$ and $|\mathbb{E}[f(Y)] + c - \mathbb{Q}(A)| < \epsilon/2$. Then

$$\begin{aligned} |\mathbb{P}(A) - \mathbb{Q}(A)| &= |[\mathbb{E}[f(Y)] + c - \mathbb{Q}(A)] - [\mathbb{E}[f(X)] + c - \mathbb{P}(A)]| \\ &\leq |\mathbb{E}[f(Y)] + c - \mathbb{Q}(A)| + |\mathbb{E}[f(Y)] + c - \mathbb{Q}(A)| \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

Since, A and ϵ were arbitrary, we must have that $\mathbb{P}(A) = \mathbb{Q}(A)$ for all $A \in \mathcal{B}$ so that $\mathbb{P} = \mathbb{Q}$. This shows that μ is injective so that k is characteristic.

This next theorem connects characteristic functions (Fourier transforms) with characteristic kernels. We say that a kernel, k, is *translation invariant* if there is another function ϕ such that for all $x, y \in \mathcal{X}, k(x, y) = \phi(x - y)$. Recall that the Fourier transform of a function, ϕ, is defined as

$$\tilde{\phi}(\xi) := \int_{\mathcal{X}} e^{-it\xi}\phi(t)dt.$$

Theorem 3.6.2. *Assume that k is translation invariant. If for all $\xi \in \mathbb{R}^m$, there exists τ_0 such that*

$$\int \frac{\tilde{\phi}(\tau(u+\xi))^2}{\tilde{\phi}(u)}du < \infty \tag{3.1}$$

for all $\tau > \tau_0$ then $\mathcal{H}$ is dense in $L^2(\mathbb{P})$ for any Borel probability measure $\mathbb{P}$ on $\mathbb{R}^m$.

Proof. Without loss of generality, assume that $\phi(0) = 1$. Using positive definiteness of k, we have that $|\phi(z)|^2 \leq \phi(0)^2 = 1$. Now, since the RKHS associated with k consists of the functions such that

$$\mathcal{H} = \left\{ f \in L^2(\mathbb{R}^m) : \int \frac{\left|\tilde{f}(u)\right|^2}{\tilde{\phi}(u)}du < \infty \right\},$$

where $\tilde{f}$ and $\tilde{\phi}$ are the Fourier transforms of f and ϕ.

Let $\mathbb{P}$ be an arbitrary probability measure on $\mathbb{R}^m$ and $\xi \in \mathbb{R}^m$. Notice that the Fourier transform of $e^{-i\xi^T z}\phi(z/\tau)$ is $\tilde{\phi}(\tau(u+\xi))$. From the assumption of the theorem we must have that $e^{-i\xi^T z}\phi(z/\tau) \in \mathcal{H}$ for all $\tau > \tau_0$. Since $\phi(z/\tau) \to 1$ as $\tau \to \infty$ and $e^{-i\xi^T z}\phi(z/\tau)$ is bounded by a constant, we can use the dominated convergence theorem to ensure that

$$\mathbb{E}_{\mathbb{P}}\left[e^{-i\xi^T z} - e^{-i\xi^T z}\phi(z/\tau)\right] \to 0 \text{ as } \tau \to \infty.$$

This shows that we only have to prove that the span of $\mathcal{A} := \left\{e^{-i\xi^T z} : \xi \in \mathbb{R}^m\right\}$ is dense in $L^2(\mathbb{P})$.

Now, let $f \in L^2(\mathbb{P})$. We can assume that f is differentiable with compact support since it's in $L^2(\mathbb{P})$. Let $\epsilon > 0$ and $M = \sup_{x\in\mathbb{R}} |f(x)|$, and A such that $[-A, A]^m$ contains the support of f and $\mathbb{P}([-A, A]^m) > 1-\epsilon/4M^2$. From a functional analysis text, we have that the series of functions

$$f_N(z) = \sum_{n_1=-N}^{N} \cdots \sum_{n_m=-N}^{N} c_n \exp\left\{\frac{i\pi}{A} n^T z\right\} \text{ for } N \in \mathbb{N}$$

where

$$c_n = \frac{1}{(2A)^m} \int_{[-A,A]^m} f(z) \exp\left\{\frac{i\pi}{A} n^T z\right\} dz$$

converges uniformly to $f(z)$ on $[-A, A]^m$ as $N \to \infty$. Then there must be some N large enough such that $|f(z) - f_N(z)|^2 < \epsilon/2$ on $[-A, A]^m$ and from the definition of $f_N(z)$ it must be the case that

$$\sup_{x\in\mathbb{R}^m} |f_N(z)|^2 < \left(M + \sqrt{\frac{\epsilon}{2}}\right) < 2M^2.$$

Then we have that $\mathbb{E}_{\mathbb{P}} |f - f_N|^2 < \epsilon$ so that span$\mathcal{A}$ is dense in $L^2(\mathbb{P})$.

Reference [34] uses a concept they coin $\mathcal{F}$-correlation defined as

$$\begin{aligned}\rho &= \max_{(f_1,f_2)\in\mathcal{F}\times\mathcal{G}} \text{corr}(\langle\Phi_1(x_1), f_1\rangle, \langle\Phi_2(x_2), f_2\rangle \\ &= \max_{(f_1,f_2)\in\mathcal{F}\times\mathcal{G}} \text{corr}\left(f_1(x_1), f_2(x_2)\right),\end{aligned}$$

where Φ_1, Φ_2 are the maps into the respective RKHS. Essentially, this measure of correlation is similar to the cross-covariance in that it is also a measure of general independence and equal to zero if and only if the random variables are independent. Reference [34] was originally written for kernel ICA but the proof makes it clear how kernel are used for independence.

Theorem 3.6.3. *Let X_1 and X_2 be random variables in $\mathcal{X} = \mathbb{R}^p$ with corresponding RKHS, $\mathcal{H}$ and kernel K. If K is characteristic, then $\rho_{\mathcal{H}} = 0$ iff X_1 and X_2 are*

independent.

Proof.

[⇒] Assume that X_1 and X_2 are independent. Then for every $f, g \in \mathcal{H}$, $f(X_1)$ is independent of $g(X_2)$ (Durrett, thm 2.1.6) so that $\text{cov}(f(X_1), g(X_2)) = 0$.

[⇐] Assume that $\rho = 0$. Then, since

$$0 = \max_{f,g \in \mathcal{H}} |\text{corr}\left(f(X_1), g(X_2)\right)|$$

it must be the case that for every $f, g \in \mathcal{H}$, $\text{cov}(f(X_1), g(X_2)) = 0$ or, equivalently, that $\mathbb{E}\left[f(X_1)g(X_2)\right] = \mathbb{E}\left[f(X_1)\right]\mathbb{E}\left[g(X_2)\right]$.

Now consider the function $\phi(x) = e^{-x^2/2\tau^2}e^{i\xi x}$ whose Fourier transform is $\tilde{\phi}(\xi) = \sqrt{2\pi}\tau e^{-\tau^2(\xi-\xi_0)^2/2}$ for $\xi_0 \in \mathbb{R}$. Further, $\tilde{\phi}$ satisfies 3.1 when $\tau > \sigma/\sqrt{2}$ so that $\phi \in \mathcal{H}$. So, for all $\xi_1, \xi_2 \in \mathbb{R}$, we have that

$$\mathbb{E}\left(e^{i\xi_1 X_1 + i\xi_2 X_2} e^{-(X_1^2+x_2^2)/2\tau^2}\right) = \mathbb{E}\left(e^{i\xi_1 X_1} e^{-X_1^2/2\tau^2}\right)\mathbb{E}\left(e^{i\xi_2 X_2} e^{-x_2^2/2\tau^2}\right).$$

As $\tau \to \infty$, we have that

$$\mathbb{E}\left(e^{i\xi_1 X_1 + i\xi_2 X_2}\right) = \mathbb{E}\left(e^{i\xi_1 X_1}\right)\mathbb{E}\left(e^{i\xi_2 X_2}\right).$$

Since, the equation above holds for all $\xi_1, \xi_2 \in \mathbb{R}$ these define characteristic functions for X_1 and X_2. And, since the joint characteristic function of (X_1, X_2) can be factored into the product of characteristics functions of independent random variables, it must be the case that X_1 and X_2 are independent since these characteristic functions uniquely determine the distributions of X_1 and X_2.

Cross covariance operators on RKHSs, in essence, define a map between RKHSs so that the inner product in the range space is the covariance of under all transformations in both RKHSs. If a kernel is characteristics, this allows us to represent

high order moments.

Let X, Y, Z be random variables in $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ with corresponding RHKSs $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}, \mathcal{H}_\mathcal{Z}$, and kernels $k_\mathcal{X}, k_\mathcal{Y}, k_\mathcal{Z}$, respectively. Assume

$$\mathbb{E}\left[k_\mathcal{X}(X,X)\right] < \infty, \mathbb{E}\left[k_\mathcal{Y}(Y,Y)\right] < \infty, \mathbb{E}\left[k_\mathcal{Z}(Z,Z)\right] < \infty.$$

This ensures that for each random variable, its RKHS is contained within its L^2 space. The *cross covariance operator* is the mapping $\Sigma_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ that satisfies

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}[f(X)g(Y)] - \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

for all $f \in \mathcal{H}_\mathcal{X}$ and $g \in \mathcal{H}_\mathcal{Y}$. It is called the *covariance operator*, Σ_{XX} if $X = Y$. The *conditional cross covariance operator* is defined as

$$\Sigma_{YX|Z} = \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}.$$

In general there is a parallel between covariance operators and covariance matrices is that the operators are a generalization of the matrices. [35] showed that there exists a unique, bounded (operator norm) operator, $V_{YX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ coined the *normalized cross covariance operator* such that

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$$

In some cases, it is useful to define Σ_{YX} in terms of its representation in the product space of $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$ which is isomorphic to the set of Hilbert-Schmidt operators (explained more in depth in the next section) from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$, $HS(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y})$, defined by the map

$$\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y} \to HS(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y})$$

$$\sum_i f_i \otimes g_i \mapsto \left[h \mapsto \sum_i \langle h, f_i \rangle_{\mathcal{H}_\mathcal{X}} g_i \right]$$

as in [36].

Using this paradigm, we can think of

$$\Sigma_{YX} = \mathbb{E}_{YX}[k_{\mathcal{Y}}(Y,\cdot) \otimes k_{\mathcal{X}}(X,\cdot)] - \mu(\mathbb{P}_Y) \otimes \mu(\mathbb{P}_X) = \mu(\mathbb{P}_{YX}) - \mu(\mathbb{P}_X \otimes \mathbb{P}_Y).$$

Changing this slightly for simplicity to the uncentered cross covariance operator,

$$\Sigma_{YX} = \mathbb{E}_{YX}[k_{\mathcal{Y}}(Y,\cdot) \otimes k_{\mathcal{X}}(X,\cdot)],$$

then

$$\begin{aligned}
\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_{\mathcal{Y}}} &= \langle \mathbb{E}_{YX}[k_{\mathcal{Y}}(Y,\cdot) \otimes k_{\mathcal{X}}(X,\cdot)], g \otimes f\rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} \\
&= \mathbb{E}_{YX}\left[\langle k_{\mathcal{Y}}(Y,\cdot) \otimes k_{\mathcal{X}}(X,\cdot), g \otimes f\rangle_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}}\right] \\
&= \mathbb{E}_{YX}\left[\langle g, k_{\mathcal{Y}}(Y,\cdot)\rangle_{\mathcal{H}_{\mathcal{Y}}} \langle f, k_{\mathcal{X}}(X,\cdot)\rangle_{\mathcal{H}_{\mathcal{X}}}\right] \\
&= \mathbb{E}_{YX}\left[g(Y) f(X)\right].
\end{aligned}$$

Further, the assumption of integrability ensures that the cross covariance operator exists.

$$\begin{aligned}
\|\Sigma_{YX}\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} &= \|\mathbb{E}_{YX}\left[k_{\mathcal{Y}}(Y,\cdot) \otimes k_{\mathcal{X}}(X,\cdot)\right]\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} \\
&\leq \mathbb{E}_{YX}\|[k_{\mathcal{Y}}(Y,\cdot) \otimes k_{\mathcal{X}}(X,\cdot)]\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}} \\
&\leq \mathbb{E}_{YX}\left[\sqrt{k_{\mathcal{X}}(X,X) k_{\mathcal{Y}}(Y,Y)}\right] \\
&\leq \sqrt{\mathbb{E}[k_{\mathcal{X}}(X,X)]\mathbb{E}[k_{\mathcal{Y}}(Y,Y)]} < \infty
\end{aligned}$$

[26] shows the following theorems which show the relationship between the cross covariance operators and independence.

Theorem 3.6.4. *1. If the product kernel $k_{\mathcal{X}} k_{\mathcal{Y}}$ is characteristic, then $V_{YX} = 0$*

if and only if $X \perp\!\!\!\perp Y$.

2. *Let* $\ddot{X} = (X, Z)$ *and* $k_{\ddot{\mathcal{X}}} = k_{\mathcal{X}} k_{\mathcal{Z}}$. *If* $k_{\ddot{\mathcal{X}}} k_{\mathcal{Y}}$ *is characteristic on* $\mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$ *and* $\mathcal{H}_{\mathcal{Z}} + \mathbb{R}$ *is dense in* $L^2(\mathbb{P}_{\mathcal{Z}})$ *then* $V_{Y\ddot{X}|Z} = 0$ *if and only if* $X \perp\!\!\!\perp Y|Z$.

Proof. Here I will prove (1). Assume that $X \perp\!\!\!\perp Y$. Using [27], theorem 2.1.6, we have that for any measurable function, f, g, $f(X) \perp\!\!\!\perp g(Y)$ therefore $\text{cov}(f(X), g(Y)) = 0$. Since this must be the case for all $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$, $\Sigma_{YX} = 0$; that is, it is the operator that sends everything in $\mathcal{H}_{\mathcal{X}}$ to the zero element in $\mathcal{H}_{\mathcal{Y}}$.

Now, if $\Sigma_{YX} = 0$, then $\mu(\mathbb{P}_X \otimes \mathbb{P}_Y) - \mu(\mathbb{P}_X)\mu(P_Y) = 0$. Since $k_{\mathcal{X}} k_{\mathcal{Y}}$ is characteristic, we must have that $\mathbb{P}_{YX} = \mathbb{P}_Y \mathbb{P}_X$ so X and Y are independent.

In order for this theory to be helpful in an empirical setting, it is necessary to be able to estimate Σ_{YX} and ideally translate it to a scalar representing its norm (so that $\Sigma_{YX} = 0$ if and only if $\|\Sigma_{YX}\| = 0$). We can do this using the Hilbert-Schmidt norm for operators. Assuming that our RHKSs are separable[1], the Hilbert-Schmidt norm of an operator,$A : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$, is

$$\|A\|_{HS}^2 = \sum_{i \in I} \sum_{j \in J} \left| \langle y_j, A x_i \rangle_{\mathcal{H}_{\mathcal{Y}}} \right|^2$$

where $\{x_i\}_i$ and $\{y_i\}_i$ are countable bases of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, respectively.

Given a independently, identically distributed sample, $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$, we can estimate Σ_{YX} using a plug-in estimate so that $\widehat{\mu}_X^{(n)} = \frac{1}{n} \sum_{i=1}^n k_{\mathcal{X}}(\cdot, X_i)$ and $\widehat{\mu}_Y^{(n)} = \frac{1}{n} \sum_{i=1}^n k_{\mathcal{Y}}(\cdot, Y_i)$. Similarly, the plug-in estimator of Σ_{YX} is

$$\widehat{\Sigma}_{YX}^{(n)} = \frac{1}{n} \sum_{i=1}^n \left(k_{\mathcal{Y}}(\cdot, Y_i) - \widehat{\mu}_Y^{(n)} \right) \left\langle k_{\mathcal{X}}(\cdot, X_i) - \widehat{\mu}_X^{(n)}, \cdot \right\rangle_{\mathcal{H}_{\mathcal{Y}}}.$$

[1] A Hilbert space is separable if it has a countable basis

Defining $\widehat{\Sigma}^{(n)}_{XX}$ and $\widehat{\Sigma}^{(n)}_{YY}$ analogously, we can estimate V_{YX} and $V_{YX|X}$ as

$$\widehat{V}^{(n)}_{YX} = \left(\widehat{\Sigma}^{(n)}_{YY} + \epsilon_n I\right)^{-1/2} \widehat{\Sigma}^{(n)}_{YX} \left(\widehat{\Sigma}^{(n)}_{XX} + \epsilon_n I\right)^{-1/2}$$

where $\epsilon_n > 0$ is a regularizing constant which guarantees inversion as shown in [34], and

$$\widehat{V}^{(n)}_{YX|Z} = \widehat{V}^{(n)}_{YX} - \widehat{V}^{(n)}_{YZ}\widehat{V}^{(n)}_{ZX}.$$

Following [26], we can use the Gram centered matrices, G_X, G_Y, G_Z, to estimate these operators as follows

$$G_{X,ij} = \left\langle k_{\mathcal{X}}\left(\cdot, X_i\right) - \widehat{\mu}^{(n)}_X, k_{\mathcal{X}}\left(\cdot, X_j\right) - \widehat{\mu}^{(n)}_X \right\rangle_{\mathcal{H}_{\mathcal{X}}};$$

however, in practice, we use $H = I - \frac{1}{n}11^T$ so center the Gram matrix,

$$G_X = H\left[k_{\mathcal{X}}(X_i, X_j)\right]_{ij} H$$

and we define G_Y and G_Z analogously. Next, define $R_X = G_X(G_X + n\epsilon I_n)^{-1}$ and R_Y and R_Z the same way as well. Using these definitions, we can construct test statistic of the Hilbert-Schmidt norm as

$$\left\|\widehat{V}^{(n)}_{YX}\right\|_{HS} = \mathrm{Tr}\left[R_Y R_X\right]$$

and

$$\left\|\widehat{V}^{(n)}_{YX|Z}\right\|_{HS} = \mathrm{Tr}\left[R_{\ddot{Y}} R_{\ddot{X}} - 2R_{\ddot{Y}} R_{\ddot{X}} R_Z + R_{\ddot{Y}} R_Z R_{\ddot{X}} R_Z\right].$$

To evaluate these measures of independence, [26] uses permutation tests for marginal independence to simulate the distribution of $\left\|\widehat{V}^{(n)}_{YX}\right\|_{HS}$ under the null hypothesis. For the conditional test, they partitioned Z_i (the conditioning random variable), into bins then permuted X_i and Y_i within those bins.

[24], building on [26], showed that, under some minor changes, that $\left\|\widehat{V}^{(n)}_{YX|Z}\right\|_{HS}$

has a Gaussian chaos distribution asymptotically:

$$p = \frac{1}{n} \sum_{k=1}^{n^2} \lambda_k z_k^2$$

where λ_k are nuisance parameters and z_k are Gaussian. In practice, this can be estimated using a gamma distribution. Having an a way to estimate the distribution under the null hypothesis is helpful especially for the conditional independence test since a permutation test may not work with real data if Z is not easily binned. In general, however, hypothesis testing capable of testing for strong conditional independence has been a hard problem [37].

Chapter 4

Mixed Conditional Mutual Information

4.1 Introduction

Estimating the dependence between random variables or vectors from data when the underlying distribution is unknown is central to statistics and machine learning. In most scientific applications, it is necessary to determine if dependence is mediated through other variables. Mutual information (MI) and conditional mutual information (CMI) are attractive for this purpose because they characterize marginal and conditional independence (they are equal to zero if and only if the variables or vectors in question are marginally or conditionally independent), and they adhere to the data processing inequality (transformations never increase information content) [38].

While there has been limited use of information theoretic statistics in specific research areas such as gene regulatory networks [39, 40, 41], this has tended to be the exception rather than the norm. Typically, it is more common to use generalized linear regression despite its inability to capture nonlinear relationships [42]. This may be, in part, because until recently, empirically estimating mutual information was only possible for exclusively discrete or exclusively continuous random variables, a severe limitation for these fields.

In this paper, we briefly review methods leading up to the estimation of MI and CMI using distribution-free, nearest-neighbors approaches. We extend the existing work to develop an estimator for MI and CMI that can handle mixed data types with improved performance over current methods. We prove that our estimator is theoretically consistent and show its performance empirically. Our research code can be found at `https://github.com/omesner/knncmi`.

4.2 Background

The MI between two random variables (or vectors) is a measure of dependence quantifying the amount of "information" shared between the random variables. The CMI

between two random variables given a third is a measure of dependence quantifying the amount of information shared between random variables given the knowledge of a third random variable or vector. These concepts were first developed by Shannon [43]; the standard modern treatment is [44]. These concepts are inherently linked to entropy and sometimes defined in terms of entropy.

4.2.1 Measure Theoretic Information

Traditionally, the information-theoretic metrics, entropy and differential entropy have been used separately for discrete and continuous random variables, respectively; however, they largely share the same properties [44]. Moreover, both of these metrics are equivalent to the expected value of a log-transformed Radon-Nikodym (RN) derivative (or density function), $\mathbb{E}\left[\log \frac{dP}{d\mu}\right]$, where P is a probability measure and μ is a reference measure. The primary distinction between entropy and differential entropy is the choice of reference measure, μ, using the Lebesgue measure, λ, for continuous random variables and the counting measure for discrete variables.

A more general construction allows for mixed probability measures with both discrete and continuous regions in their domain. Reference [45, §5.5] defines entropy (and other divergences) for generalized probability spaces as the supremum of all finite, discrete representations (quantizers) of random variables, mirroring the definition of the Lebesgue integral. Because our problem is concerned specifically with mixed, discrete-continuous measures, we use this explicit definition which is helpful when calculating theoretical values and assume all measurable spaces are standard according to [45, §1.4].

A nuance specific to this generalized framework is determining an appropriate reference measure, μ, such that P is absolutely continuous with respect to μ, $P \ll \mu$. Similarly, MI and CMI are expected values of the log-transformed RN derivative of a joint probability measure with respect to a product measure of marginal or conditional factors, requiring that the joint measure be absolutely continuous with respect

to the product. For discrete measures, this is not a concern, but mathematically critical for continuous ones due to the Radon-Nikodym theorem. The lack of absolute continuity for MI and CMI can occur when a continuous random variable is a deterministic function of another for any region of the joint space. For mixed probability measures, we avoid this for the continuous region of a probability measure's domain by characterizing cross-sections of arbitrary subsets.

Definition 4.2.1. *Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}, P_{XY})$ be a probability space with marginal probability measures, P_X and P_Y. P_{XY} is **non-singular** if for any measurable set, $E \subseteq \mathcal{X} \times \mathcal{Y}$, $a \in \mathcal{X}$ and $b \in \mathcal{Y}$, such that $P_X(E_b) = 0$ and $P_Y(E_a) = 0$, then $P_{XY}(E) = 0$ where $E_b = \{x : (x, b) \in E\}$ and $E_a = \{y : (a, y) \in E\}$.*

Notice that discrete measures are vacuously non-singular, but continuous measures will be singular if supported on a Lebesgue-measure zero subset of the joint space.

Requiring that a mixed, joint probability measure is non-singular ensures the existence of a product reference measure:

Lemma 4.2.1. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional probability space such that for each $i \in [d] := \{1, 2, \ldots, d\}$, there exists a $\mathcal{B}_i$-measurable set $E \subseteq \mathcal{X}_i$ such that $P_i \ll \lambda$ on E and $\lambda(\mathcal{X} \backslash E) = 0$. If P is non-singular then there exists a d-dimensional product measure $\mu = \prod_{i=1}^{d} \mu_i$ on $\mathcal{X}$ such that $P \ll \mu$ such that for each $i \in [d]$, $\mu_i = \lambda + I_C$ where $C = \{x \in \mathcal{X}_i : P_i(\{x\}) > 0\}$.*

Further, if all conditional probability measures are regular, non-singularity ensures the existence of MI and CMI as shown in the following theorem.

Theorem 4.2.2. *Let P_{XYZ} be a joint probability measure on the space $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, where $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are all metric spaces. If for every value of Z, $P_{XY|Z}$ is non-singular (see def. 4.2.1), then $\frac{dP_{XY|Z}}{d\left(P_{X|Z} \times P_{Y|Z}\right)}$ is well-defined.*

Proof. For the RN derivative to exist, the Radon-Nikodym theorem requires that $P_{XY|Z}$ is absolutely continuous with respect to $P_{X|Z} \times P_{Y|Z}$, $P_{XY|Z} \ll P_{X|Z} \times$

$P_{Y|Z}$. Because we assume that all conditional probabilities are regular, we omit the argument associated with Z and proceed as probabilities measures of X and Y as appropriate.

Assume $A \subseteq \mathcal{X}\times\mathcal{Y}$ such that $(P_{X|Z}\times P_{Y|Z})(A) = 0$. Define $A_1 = \left\{x : P_{Y|Z}(A_x) > 0\right\}\times \mathcal{Y}$, $A_2 = \mathcal{X} \times \left\{y : P_{X|Z}(A_y) > 0\right\}$, and $A_3 = \left\{(x, y) : P_{X|Z}(A_y) = P_{Y|Z}(A_x) = 0\right\}$. Notice that $A \subseteq A_1 \cup A_2 \cup A_3$.

From Fubini's theorem, we have that

$$\begin{aligned} 0 &= (P_{X|Z} \times P_{Y|Z})(A) \\ &= \int_{\mathcal{X}} P_{Y|Z}(A_x) dP_{X|Z}(x). \end{aligned}$$

Using [46, Lemma 1.3.8], $f \geq 0, \int f d\mu = 0 \Rightarrow \mu\{x : f(x) > 0\} = 0$, for the first equality, we must have

$$\begin{aligned} 0 &= P_{X|Z}\left(\left\{x : P_{Y|Z}(A_x) > 0\right\}\right) \\ &= P_{XY|Z}\left(\left\{x : P_{Y|Z}(A_x) > 0\right\} \times \mathcal{Y}\right) \\ &= P_{XY|Z}\left(A_1\right). \end{aligned}$$

Using the same construction but switching X and Y, we also have that $0 = P_{XY|Z}\left(A_2\right)$. $P_{XY|Z}\left(A_3\right) = 0$ follows from the definition of nonsingular.

This shows that $P_{XY|Z} \ll P_{X|Z} \times P_{Y|Z}$. Now, we may apply the RN theorem, so there exists a measurable function, f such that for any measurable set $A \subseteq \mathcal{X} \times \mathcal{Y}$,

$$\int_A f d(P_{X|Z} \times P_{Y|Z}) = P_{XY|Z}(A) \tag{4.1}$$

and f is unique almost everywhere $P_{X|Z} \times P_{Y|Z}$. □

[45, Lemmas 7.16 and 7.17] show that if a joint measure is absolutely continuous with respect to any product measure, then it is absolutely continuous with respect to its product measure. Theorem 4.2.2 maybe more helpful for data analysis by

showing the sufficient condition for a nonsingular distribution in the mixed setting in def. 4.2.1.

Definition 4.2.2. *The **conditional mutual information** of X and Y given Z is*

$$I(X;Y|Z) \equiv \int \log \left(\frac{dP_{XY|Z}}{d\left(P_{X|Z} \times P_{Y|Z}\right)} \right) dP_{XYZ} \tag{4.2}$$

*where $P_{XY|Z}, P_{X|Z}$, and $P_{Y|Z}$ are regular conditional probability measures and $\frac{dP_{XY|Z}}{d\left(P_{X|Z} \times P_{Y|Z}\right)}$ is the Radon-Nikodym derivative of the joint conditional measure, $P_{XY|Z}$, with respect to the product of the marginal conditional measures, $P_{X|Z} \times P_{Y|Z}$. If Z is constant, then Eq. 4.2 is $I(X;Y)$, the **mutual information** of X and Y.*

Definition 4.2.2 retains the standard properties of CMI.

Corollary 1. *1. X and Y are conditionally independent given Z, $X \perp\!\!\!\perp Y|Z$, if and only if $I(X:Y|Z) = 0$.*

2. $I(X;Y|Z) = H(X,Y,Z) - H(X,Z) - H(Y,Z) + H(Z)$

3. If $X \to Z \to Y$ is a Markov chain, the data processing inequality states that $I(X;Y) \leq I(X;Z)$.

4.2.2 Estimators for Discrete Random Variables

Many estimators of entropy, MI, and CMI for discrete random variables are based on straight-forward "plug-in" estimates, substituting the empirical distribution in to the defining formulas. The entropy plug-in estimator for discrete random variables with a finite range (alphabet),

$$\widehat{H}(X) := -\sum_{X=i} \widehat{p}_i \log \widehat{p}_i, \tag{4.3}$$

where $\widehat{p}_i := \mathbb{P}(X = i)$ is estimated from the data, is asymptotically Gaussian [47]. More generally, information theoretic plug-in estimators for random variables with

countable ranges are universally consistent under some conditions and without these conditions, can converge arbitrarily slowly [48]. However, such estimates can suffer from substantial finite-sample bias, especially when the number of categories is large [49].

In cases where the range is large, relative to the number of samples, replacing the logarithm in the entropy plug-in with the digamma function, $\psi(x) := \frac{d}{dx} \log \Gamma(x)$, may decrease bias [50]. Within the Bayesian estimation paradigm, the use of the digamma function emerges from the Dirichlet distribution as the conjugate prior for the multinomial distribution [51]. Other, problem-specific estimators have emerged as well, such as neural spike trains [52] and ecology [53], for example.

4.2.3 Estimators for Continuous Random Variables

Estimation for continuous random variables is also challenging. A direct plug-in estimation would first require density estimates, which is a challenging problem in itself. Dmitriev and Tarasenko first proposed such an estimator for functionals [54] for scalar random variables. Darbellay and Vajda [55], in contrast, proposed an estimator mutual information based on frequencies in rectangular partitions. Nearest-neighbor methods of estimating information-theoretic quantities for continuous random variables which evade the step of directly estimating a density go back over thirty years, to Kozachenko and Leonenko [56], which proposed an estimator of the differential entropy.

Kozachenko and Leonenko estimator of entropy

Kozachenko and Leonenko (KL) first used nearest neighbors to estimate differential entropy [56]. Briefly, let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ be a random variable and $x_1, \ldots, x_n \sim P_X$ be a random sample from X. Estimating the entropy of X, as

$$\widehat{H}(X) = -\frac{1}{n} \sum_{i=1}^{n} \log \widehat{f_X(x_i)} \tag{4.4}$$

where f_X is the density of X, we focus on $\widehat{\log f_X(x_i)}$ for each i locally. Define $\rho_{k,i,p}$ as the ℓ_p-distance from point x_i to its kth nearest neighbor, $k\text{NN}_i$, and $B(x_i, \rho_{k,i,p})$ as the d-dimensional, ℓ_p ball of radius $\rho_{k,i,p}$ centered at x_i. Consider the probability mass of $B(x_i, \rho_{k,i,p})$, $P_{k,i,p} \equiv P_X(B(x_i, \rho_{k,i,p}))$. $P_{k,i,p}$ could be estimated using the d-dimensional volume in ℓ_p of $B(x_i, \rho_{k,i,p})$ [57] as

$$P_{k,i,p} \approx f_X(x_i) c_{d,p} \rho_{k,i,p}^d \tag{4.5}$$

where $c_{d,p} = 2^d \Gamma\left(1 + \frac{1}{p}\right)^d / \Gamma\left(1 + \frac{d}{p}\right)$ if $f_X(x_i)$ were known. Notice that, intuitively, $P_{k,i,p} \approx \frac{k}{n}$. In fact, using lemma 4.6.6 and seeing that the integral is the same as $\mathbb{E}[\log V]$ for $V \sim \text{Beta}(k, n-k)$,

$$\mathbb{E}[\log P_{k,i,p}] = \psi(k) - \psi(n) \tag{4.6}$$

where $\psi(x) = d \log \Gamma(x)/dx$ is the digamma function, and does not depend on choice of p. Substituting the estimate for $P_{k,i,p}$ in approximation (4.5) into the expectation in (4.6), we have the estimator for $\widehat{\log f_X(x_i)}$:

$$\widehat{\log f_X(x_i)} = \psi(k) - \psi(n) - \log c_{d,p} - d \log \rho_{k,i,p}, \tag{4.7}$$

making the KL estimator

$$\widehat{H}(X) = -\psi(k) + \psi(n) + \log c_{d,p} + \frac{d}{n} \sum_{i=1}^{n} \log \rho_{k,i,p}. \tag{4.8}$$

[58] showed that its bias is $\tilde{O}(n^{-1/d})$ and variance is $\tilde{O}(1/n)$ where $\tilde{O}$ is the limiting behavior up to polylogarithmic factors in n.

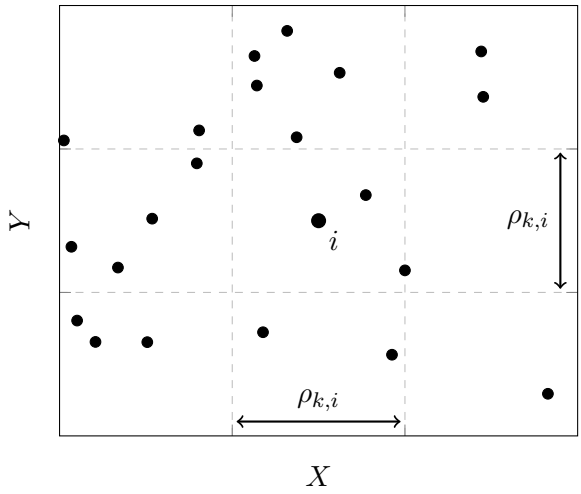


Figure 4.1: The scatter plot above shows point i and its kNN where $k = 2$ on the right vertical dashed line. Here $n_{X,i} = 9$ and $n_{Y,i} = 6$.

Kraskov, Stögbaur, and Grassberger estimator of mutual information

Kraskov, Stögbaur, and Grassberger (KSG) [59] developed an estimator for MI based on $I(X,Y) = H(X) + H(Y) - H(X,Y)$ and a variation of the KL entropy estimator for continuous random variables or vectors, X and Y in $\mathbb{R}^{d_X}$ and $\mathbb{R}^{d_Y}$, respectively. Let $(x_1, y_1), \ldots, (x_n, y_n) \sim P_{XY}$. Setting $p = \infty$, define the ℓ_∞-distance from point (x_i, y_i) to its kNN as $\frac{1}{2}\rho_{k,i,\infty} \equiv \frac{1}{2}\rho_{k,i}$, so that $c_{d,p} = 1$ and $\log c_{d,p} = 0$. Using this, the local KL estimate for the (negative) joint entropy at point i is

$$\widehat{\log f_{XY}}_i = -\psi(k) + \psi(n) + (d_X + d_Y) \log \rho_{k,i}; \tag{4.9}$$

$\widehat{H}(X,Y)$ is computed as in eq. (4.4). To calculate $\widehat{\log f_X}_i$ and $\widehat{\log f_Y}_i$, the KSG method deviates slightly from KL by using different values for the k hyper-parameter argument for each i. In contrast, $\widehat{\log f_{XY}}_i$ used the same value of k for each i to calculate $\widehat{H}(X,Y)$. For KL, the k argument can be chosen as any integer value between 1 and $n-1$ which in turn determines the ℓ_p-distance to each point's kNN. Considering each point i separately, KSG works backward for $\widehat{H}(X)$ and $\widehat{H}(Y)$, by

first choosing a distance, r, then counting the number of points that fall within the ℓ_∞ ball of radius r centered at point i within the X (or Y) subspace. It uses this count of points to compute $\widehat{\log f_{X\,i}}$ (or $\widehat{\log f_{Y\,i}}$) in place of k hyper-parameter argument and r in the distance argument. Specifically, for each i, KSG chooses $r = \frac{1}{2}\rho_{k,i}$, the ℓ_∞-distance from point (x_i, y_i) to its kNN in $(\mathbb{R}^{d_X+d_Y}, \ell_\infty)$, that was used to calculate $\widehat{\log f_{XY\,i}}$. Call the corresponding count of points $n^*_{X,i}$ in the X subspace and $n^*_{Y,i}$ in the Y subspace:

$$n^*_{W,i} = \left| \left\{ w_j : \|w_i - w_j\|_\infty < \frac{1}{2}\rho_{k,i}, i \neq j \right\} \right| \tag{4.10}$$

and

$$\widehat{\log f_{W\,i}} = -\psi(n^*_{W,i}) + \psi(n) + d_W \log \rho_{k,i} \tag{4.11}$$

where W is either X or Y.

The KL estimator is accurate because the value of $P_{k,i,p}$, the probability in the local neighborhood around (x_i, y_i) extending out to its kNN, is completely determined by k and n. By using the ℓ_∞ norm in the KSG estimator, $\frac{1}{2}\rho_{k,i}$ is equal to the absolute scalar difference between point (x_i, y_i) and $k\text{NN}_i$ at a coordinate in either X or Y. This way, the entropy estimates for either X or Y will be accurate in the KL paradigm. But, the point $k\text{NN}_i$ is not counted in $n^*_{X,i}$ or $n^*_{Y,i}$ because the definition counts points whose distance from (x_i, y_i) are strictly less than $\frac{1}{2}\rho_{k,i}$, biasing (4.7) toward zero. Using $n^*_{W,i} + 1$ for $W = X, Y$ corrects this for either X or Y but not both; that is, either $n_{X,i} + 1$ or $n_{Y,i} + 1$ will be the number of points within a distance of exactly $\frac{1}{2}\rho_{k,i}$. See Fig. 4.1.

Plugging the estimates for $\widehat{H}(X, Y), \widehat{H}(X)$ and $\widehat{H}(Y)$ discussed above into $I(X, Y) = H(X, Y) - H(X) - H(Y)$, we have

$$\widehat{I}_{\text{KSG}}(X; Y) = \psi(k) + \psi(n) - \frac{1}{n}\sum_{i=1}^{n} \left[\psi(n^*_{X,i} + 1) + \psi(n^*_{Y,i} + 1)\right] \tag{4.12}$$

where the $\frac{d}{n}\sum_{i=1}^{n}\rho_{k,i}$ terms all cancel from using the same value of $\rho_{k,i}$ for each i and $\log c_{d,p}$ is zero using the ℓ_∞ norm and choosing to set the kNN distance to $\frac{1}{2}\rho_{k,i}$.

The original work [59] did not offer any proofs on convergence. Attempting to correct the counting error mentioned, [59] provided another, less-used estimator as well. Gao et al. [58] later showed that the KSG estimator is consistent with a bias of $\tilde{O}\left(n^{-\frac{1}{d_X+d_Y}}\right)$ and a variance of $\tilde{O}\left(1/n\right)$. Reference [60] provides a very clear analysis assumptions and corresponding convergence rates for both the KL entropy estimator and the KSG MI estimator.

Frenzel and Pompe estimator of conditional mutual information

Using a similar technique to estimate conditional mutual information, Frenzel and Pompe (FP) first, though several other papers as well [61, 62, 63, 64, 65] used $I(X;Y|Z) = H(X,Y,Z) - H(X,Z) - H(Y,Z) + H(Z)$ combined with the KSG technique to cancel out the $\rho_{k,i}$ term from each of the entropy estimators to estimate CMI as

$$\xi_i = \psi(k) - \psi(n^*_{XZ,i}+1) - \psi(n^*_{YZ,i}+1) + \psi(n^*_{Z,i}+1) \tag{4.13}$$

where $n_{W,i}$ is calculated as in equation 4.10 with $W = XZ, YZ, Z$. The global CMI estimator, $\widehat{I}_{\mathrm{FP}}(X;Y|Z)$, is calculated by averaging overall ξ_i. While these papers show that this estimator does well empirically, they do not provide theoretical justification.

4.2.4 Estimation for Mixed Variables

Gao, Kannan, Oh, and Viswanth estimator of mutual information

Gao, Kannan, Oh, and Viswanth (GKOV) [66] expanded on the KSG technique to develop an MI estimator for mixes of discrete and continuous, real-valued random variables. In this setting, unlike the purely continuous one, there is some probability that multiple independent observations will be equal. Depending on the value of k,

there is a corresponding, nonzero probability that the kNN distance is zero for some points. While this impairs the KL entropy estimator due to the $\log \rho$ term, the KSG estimator only uses the kNN distance for counting points with that radius. Similar to the insight for the KSG technique, [66] allows k to change for points whose kNN distance is zero:

$$\tilde{k}_i^* = \left|\left\{(x_j, y_j) : \|(x_i, y_i) - (x_j, y_j)\|_\infty = 0, i \neq j\right\}\right|. \tag{4.14}$$

To accommodate points whose kNN distance is zero, [66] changes the definition of $n_{W,i}^*$ to include boundary points:

$$n_{W,i} = \left|\left\{w_j : \|w_i - w_j\|_\infty \leq \frac{1}{2}\rho_{k,i}, i \neq j\right\}\right| \tag{4.15}$$

where $\frac{1}{2}\rho_{k,i}$ remains the ℓ_∞ distance from point (x_i, y_i) to $k\text{NN}_i$. For index i, [66] locally estimates MI as

$$\xi_i = \psi(\tilde{k}_i^*) + \log(n) - \log(n_{X,i} + 1) - \log(n_{Y,i} + 1). \tag{4.16}$$

The global MI estimate is the average of the local estimates for each point:

$$\widehat{I}_{\text{GKOV}}(X;Y) = \frac{1}{n}\sum_{i=1}^{n} \xi_i \ . \tag{4.17}$$

[66] shows that this estimator is consistent under some mild assumptions.

Rahimzamani, Asnani, Viswanath, and Kannan (RAVK) [67] extend the idea of [66] for estimating MI for mixed data to a concept the authors define as *graph divergence measure*, a generalized Kullback-Leibler (KL) divergence between a joint probability measure and a factorization of the joint probability measure. The authors say that this can be thought of as a metric of incompatibility between the joint probability and the factorization.

Setting the factorization of P_{XYZ} to $P_{X|Z}P_{Y|Z}P_Z$ gives an equivalent definition

of 4.2.2 of conditional mutual information. Using this factorization, the GKOV estimator for CMI at index i is

$$\xi_i = \psi(\tilde{k}_i) - \log(n_{XZ,i} + 1) - \log(n_{YZ,i} + 1) + \log(n_{Z,i} + 1). \tag{4.18}$$

The authors state that $\tilde{k}_i$ is the number of points within, $\rho_{k,i}$, the distance to the kNN, of observation i. Giving more detail, case III in the proof for [67, theorem 2] states that $\rho_{k,i} > 0$ implies that $\tilde{k}_i = k$, suggesting that $\tilde{k}_i$ is defined the same as (4.14). Similarly, the proofs suggest that $n_{W,i}$ is defined as (4.15). The global CMI, $I_{\text{GKOV}}(X;Y|Z)$ is calculated by averaging over all ξ_i. This paper shows that this estimator is consistent with similar assumptions to those found in [66].

4.3 Proposed Information Estimators

The estimator for CMI (and MI) proposed in this paper builds on the ideas in the previous papers but with critical changes that improve performance. Start by considering local CMI estimates as in (4.13). As discussed in § 4.2.3, for index i, each negative local entropy estimate (4.7), $\widehat{\log f_{XYZ}}_i, \widehat{\log f_{XZ}}_i, \widehat{\log f_{YZ}}_i$, and $\widehat{\log f_Z}_i$, (before terms cancel) is identical to the KL paradigm when the distance from (x_i, y_i, z_i) to its kNN, $n_{XZ,i}$NN, $n_{YZ,i}$NN, and $n_{Z,i}$NN for each respective subspace (XYZ, XZ, YZ or Z) is exactly $\frac{1}{2}\rho_{k,i}$. Moving from exclusively continuous data, where ties occur with probability zero, to mixed data where ties occur with nonzero probability, required that $n^*_{W,i}$ from (4.10) include boundary points as in $n_{W,i}$ from (4.15). With this change, the entropy estimates are frequently accurate using $n_{W,i}$ rather than $n_{W,i} + 1$ for $W = XZ, YZ, Z$ for continuous data.

With the ℓ_∞ norm, the kNN distance value, $\frac{1}{2}\rho_{k,i}$, is equal to the scalar distance in at least one coordinate of the random vector (X, Y, Z). If this coordinate is in Z, then the distance term in each local entropy estimate from (4.7) will be exactly $d \log \rho_{k,i}$ for $\widehat{\log f_{XYZ}}_i, \widehat{\log f_{XZ}}_i, \widehat{\log f_{YZ}}_i$, and $\widehat{\log f_Z}_i$ (because each contains Z).

Again, this is because within each given subspace, $\rho_{k,i} = \rho_{n_{W,i},i}$, for $W = XZ, YZ, Z$, and thus

$$\begin{aligned}\xi_i &= -\widehat{\log f_{XYZ}}_i + \widehat{\log f_{XZ}}_i + \widehat{\log f_{YZ}}_i - \widehat{\log f_Z}_i \\ &= \psi(k) - \psi(n_{XZ,i}) - \psi(n_{YZ,i}) + \psi(n_{Z,i})\end{aligned}$$

with perfect cancellations.

If the ℓ_∞-distance coordinate is in X, then $\rho_{k,i} = \rho_{n_{XZ,i},i}$, so that the corresponding terms in $\widehat{\log f_{XYZ}}_i$ and $\widehat{\log f_{XZ}}_i$ cancel but the other two distance terms may not. An analogous argument can be made for Y. If the dimension of Z is greater than X and Y, heuristically, one might expect the kNN distance to fall in the larger Z dimension.

Theorem 4.3.3 shows that the proposed estimator tends to zero as the dimension of the Z vector increases if the sample size remains the same. The methods discussed in § 4.2.4 will also converge to zero as the dimension increases; however, the proposed method is an improvement, especially for discrete points. The combined dimension of (X, Y, Z) can affect the value of $\tilde{k}_i$, on discrete data, when the kNN distance is greater than zero. Consider the case where data is comprised of exclusively discrete random variables, that is; each point in the sample has a positive probability point mass. As the dimension of (X, Y, Z) grows, probability point masses will diminish as long as the added variables are not determined given the previous variable. Moreover, point masses in higher-dimensional spaces will necessarily be less than or equal to their corresponding locations in lower-dimensional spaces. It is possible that the kNN distance for index i is zero, especially if it has a large probability point mass relative to n. But, if its point mass in the XYZ-space is not sufficiently large to expect more than one point at its location for the given sample size, n, we would expect its kNN distance to be greater than zero. If the $\tilde{k}_i = k$, as it would in eq. (4.14) because $\frac{1}{2}\rho_{k,i} > 0$, then $n_{XZ,i}, n_{YZ,i}$, and $n_{Z,i}$ will be the total number of points within the distance to the kNN including

points on the boundary for the appropriate subspace, XZ, YZ or Z. But, because the data are discrete, it is possible/likely that the kNN is not unique. This would indicate that there are more than k, points at and within the same radius, $\frac{1}{2}\rho_{k,i}$ in the XYZ-space. Under counting here would bias the local estimate of CMI (4.18) downward because k would be small relative to the values $n_{XZ,i}, n_{YZ,i}$, and $n_{Z,i}$, in the associated subspaces. To fix this, we set $\tilde{k}_i$ to the number of points that are less than or equal to the kNN distance from point (x_i, y_i, z_i) as

$$\tilde{k}_i = \left| \left\{ w_j : \|w_i - w_j\| \leq \frac{1}{2}\rho_{k,i}, i \neq j \right\} \right|. \tag{4.19}$$

Notice that if the data are all continuous, then $\tilde{k}_i = k$ with probability one so this change will only affect discrete points.

Moving to the use of the digamma function, ψ, verses the natural logarithm, log, the overview of discrete estimators given in § 4.2.2, indicate that the use of digamma should improve performance, particularly if there are many discrete categories. The methods presented in § 4.2.3 to estimate MI and CMI for continuous data also indicate using the digamma function over log. In contrast, the methods in §4.2.4 use both when estimating MI and CMI for mixed data. Though no explicit reason is given for the deviation, it seems innocuous given that $|\log(w) - \psi(w)| \leq \frac{1}{w}$ for $w > 0$, and, possibly reasonable given that the plug-in estimator of CMI on discrete data is $\log(\tilde{k}_i) - \log(n_{XZ,i}) - \log(n_{YZ,i}) + \log(n_{Z,i})$ similar to (4.13), with the difference being that it uses log in place of digamma. But, the use of digamma has emerged in both the discrete and continuous literature. Moreover, using a single functional form rather than one for the discrete case and another for the continuous case, makes for a simpler estimator. For this reason, we use ψ for continuous and discrete data.

If a variable/coordinate of (X, Y, Z) is categorical (non-numeric), we use the discrete distance metric for that coordinate in the random vector, in place of absolute difference: the coordinate distance is zero at that coordinate for two observations when equal and one otherwise. Several cited lemmas and theorems used in the

proofs in § 4.3.1 assume vectors to be in $\mathbb{R}^d$. Categorical variables do not strictly satisfy this requirement but transforming categorical variables to dummy indicators (as one does in regression) yields an isometry between the categorical space with the discrete metric and $\mathbb{R}^m$ where the variable takes $m+1$ distinct values with an ℓ_∞ metric. While it is not necessary to create dummy variables for the code to work, we can be assured that the proofs are satisfied even when data include categorical data.

To calculate the proposed local CMI estimate for index i, notice that $\tilde{k}_i = k$ when the region surrounding observation i (to its kNN) is continuous but $\tilde{k}_i > k$ when discrete. Thus, using the value of $\tilde{k}_i$ is appropriate in both cases. After calculating $\tilde{k}_i$ and its kNN distance, $\frac{1}{2}\rho_{k,i}$ using eq. (4.19), $\frac{1}{2}\rho_{k,i}$ is reused to determine $n_{W,i}$ for $W \in \{XZ, YZ, Z\}$ from eq. (4.15). For each $i \in \{1, \ldots, n\}$, define

$$\xi_i = \psi\left(\tilde{k}_i\right) - \psi(n_{XZ,i}) - \psi(n_{YZ,i}) + \psi(n_{Z,i}) \ . \tag{4.20}$$

The sample estimate for the proposed CMI estimator is

$$\widehat{I}_{\text{prop}}(X;Y|Z) = \max\left\{\frac{1}{n}\sum_{i=1}^{n}\xi_i, 0\right\}. \tag{4.21}$$

To calculate MI between X and Y, we can make Z constant according to def. 4.2.2 so that $n_{Z,i} = n$. We define CMI and MI as the positive part of the mean because CMI and MI are provably non-negative. This setting can easily be changed in the code with a function argument. In the simulations shown in § 4.4, we display the mean itself for greater visibility.

4.3.1 Consistency

The estimator proposed in §4.3 is consistent for fixed-dimensional random vectors under mild assumptions. Theorem 4.3.1 shows that the estimator is asymptotically unbiased and theorem 4.3.2 shows that its asymptotic variance is zero.

As shorthand notation, we set $f \equiv \frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})}$ and for a random variable W on $\mathcal{W}$ with probability measure, P_W, and $w \in \mathcal{W}$, define

$$P_W(r) = P_W\left(\{v \in \mathcal{W} : \|v - w\|_\infty \leq r\}\right). \tag{4.22}$$

Theorem 4.3.1. *Let $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$ be an i.i.d. random sample from P_{XYZ}. Assume the following:*

1. *$k = k_n \to \infty$ and $\frac{k_n}{n} \to 0$ as $n \to \infty$.*
2. *For some $C > 0$, $f(x, y, z) \leq C$ for all $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.*
3. *$\{(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} : P_{XYZ}((x, y, z)) > 0\}$ is countable and nowhere dense in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$*

then

$$\lim_{n\to\infty} \mathbb{E}\left[\widehat{I}_{prop}(X, Y|Z)\right] = I(X, Y|Z)\ . \tag{4.23}$$

The proof can be found in App. 4.6.1.

Theorem 4.3.2. *Let $W = \{W_1, \ldots, W_n\}$ be a random samples of size n such that for each i, $W_i = (X_i, Y_i, Z_i)$, $k \geq 2$, and let $\widehat{I}_n(W) = \frac{1}{n}\sum_{i=1}^n \xi_i(W)$ where $\xi_i(W) = \xi_i$ as defined above. Then*

$$\lim_{n\to\infty} Var\left(\widehat{I}_{prop}(W)\right) = 0\ . \tag{4.24}$$

The proof can be found in App. 4.6.2.

Corollary 2. *Let $W_1, \ldots, W_n$ be an independent, identically distributed sample. Then*

$$\begin{aligned} &P\left(\left|\widehat{I}_n(W) - \mathbb{E}\left[\widehat{I}_n(W)\right]\right| > t\right) \\ &\quad \leq 2\exp\left\{\frac{-t^2 n}{2592 k^2 \gamma_d^2 (\log n)^2}\right\} \end{aligned} \tag{4.25}$$

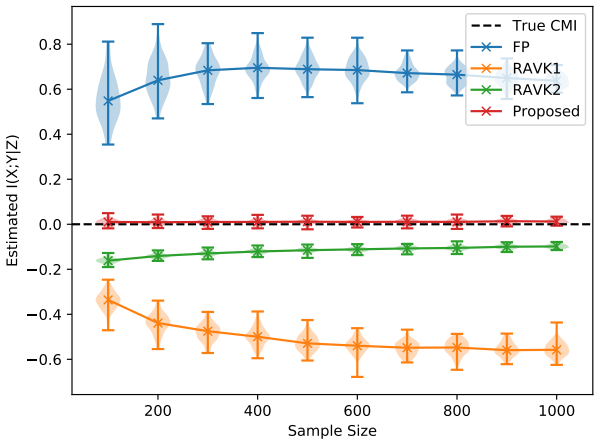


Figure 4.2: $X \sim \text{Exponential}(10)$, $Z \sim \text{Poisson}(X)$, and $Y \sim \text{Binomial}(Z, 0.5)$, $I(X;Y|Z) = 0$.

where γ_d is a constant that is only dependent on the dimension of W.

The proof can be found in App. 4.6.3.

Despite the estimator's unbiasedness in large samples, it is biased toward zero on high-dimensional data with a fixed sample size, suffering from the curse of dimensionality as kNN regression does.

Theorem 4.3.3. *Assume X and Y have fixed-dimension and that $Z = (Z_1, Z_2, \ldots, Z_d)$ is a d-dimensional random vector. If the entropy rate of Z is nonzero, that is, $\lim_{d\to\infty} \frac{1}{d} H(Z) \neq 0$, then $\widehat{I}_{prop}(X, Y|Z) \xrightarrow{P} 0$ (converges in probability) as $d \to \infty$.*

The proof can be found in App. 4.6.4.

4.4 Experiments

To evaluate the empirical performance of the proposed estimator, we compared it to the FP estimator for continuous variables found in § 4.2.3 and to two versions of the RAVK estimator for CMI in § 4.2.4 on simulated mixed data from various setting. Both RAVK1 and RAVK2 are calculated using eq (4.18), but using different values

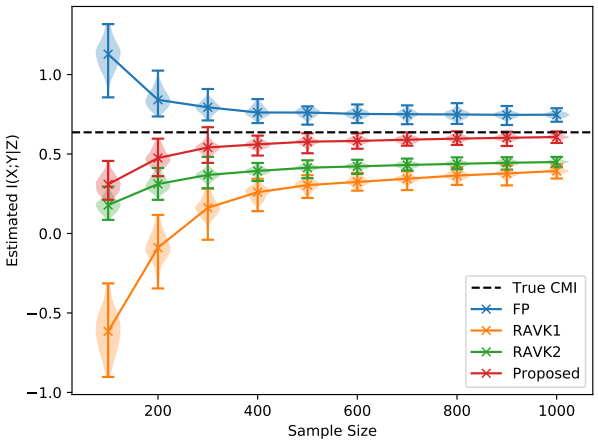


Figure 4.3: $X \sim$ Discrete Uniform$(0, 3)$ and $Y \sim$ Continuous Uniform$(X, X + 2)$, $Z \sim$ Binomial$(3, 0.5)$, $I(X; Y|Z) = \log 3 - 2 \log 2/2$.

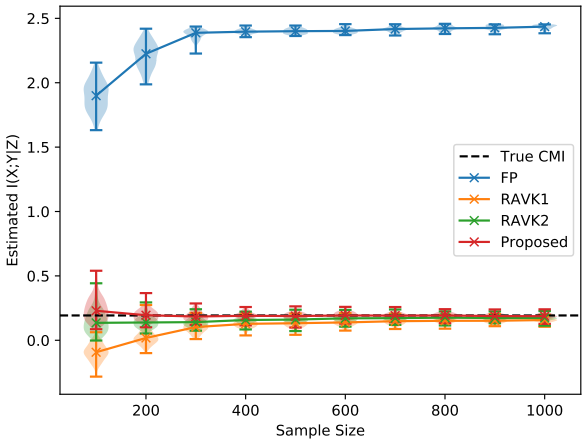


Figure 4.4: $\mathbb{P}((X, Y) = (1, 1)) = \mathbb{P}((X, Y) = (-1, -1)) = 0.4$, $\mathbb{P}((X, Y) = (1, -1)) = \mathbb{P}((X, Y) = (1, -1)) = 0.1$, $Z \sim$ Poisson(2), $I(X; Y|Z) = 2 \cdot 0.4 \log(0.4/0.5^2) + 2 \cdot 0.1 \log(0.1/0.5^2)$.

for $\tilde{k}_i$. RAVK1 uses eq (4.14) and RAVK2 uses eq (4.19). The FP estimator, as it was designed for exclusively continuous data, when $\rho_{k,i} = 0$ will compute 0 for $n^*_{w,i}$ from eq 4.10, so ψ will be undefined. In the simulations, we used $\max\left\{n^*_{w,i}, 1\right\}$. For greater visibility, all figures show positive and negative estimator values (even

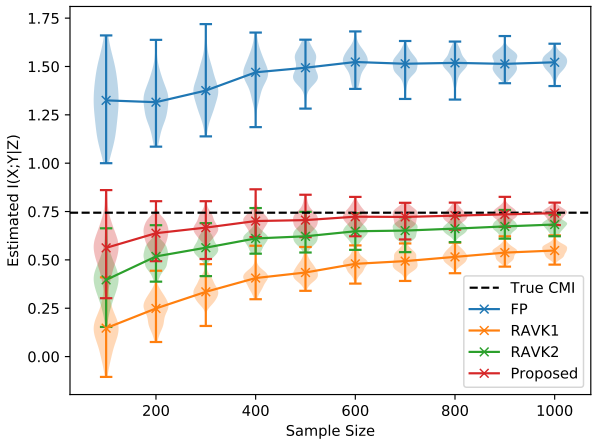


Figure 4.5: X and Y are a mixture distribution where with probability $\frac{1}{2}$, (X, Y) is multivariate Gaussian with a correlation coefficient of 0.8 and with probability $\frac{1}{2}$, (X, Y) places probability mass of 0.4 at $(1, 1)$, $(-1, -1)$ and probability mass of 0.1 at $(1, -1)$ and $(-1, 1)$, as in the third experiment. Z is an independently generated Binomial$(3, 0.2)$. $I(X;Y|Z) = 0.4 \log(2 \cdot 0.4/0.5^2) + 0.1 \log(2 \cdot 0.1/0.5^2) + 0.125 \log(4/(1 - 0.8^2))$

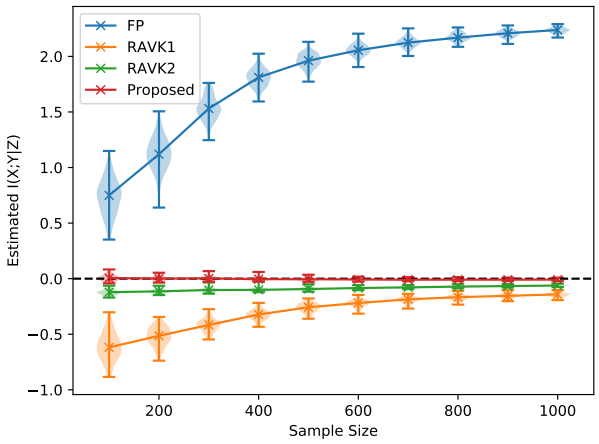


Figure 4.6: Information between race and recidivism conditioning on COMPAS recidivism prediction score. Each sample was bootstrapped for the required size. Here the true CMI is not known. The data can be found at `https://github.com/propublica/compas-analysis`

though CMI is non-negative). Specifically, all figures show the proposed estimator as $\frac{1}{n}\sum_{i=1}^{n}\xi_i$ rather than equation (4.21). All simulation data, methods code, and visuals were done in Python 3.6.5. Both the estimation package and simulation code can be found at *https://github.com/omesner/knncmi*. We simulated data from differing distributions with 100 observations up to 1000 in intervals of 100. The violin plots in Fig. 4.2–4.5 show the distribution of estimates from 100 simulated datasets for each sample size. The "$\times$" markers in each violin plot indicates the mean of all estimates and the "$-$" represent to most extreme values. For both the proposed and continuous estimator, we used $k = 7$ for all datasets.

The first simulation (Fig. 4.2) was inspired by [68, example 4.4.5]. In this scenario, a mother insect lays eggs at a random rate, $X \sim$ Exponential(10). The number of eggs she lays is $Z \sim$ Poisson(X), and the number of the eggs that survive is $Y \sim$ Binomial($Z, 0.5$). In this Markov chain ($X \to Z \to Y$), X and Y are marginally dependent $X \not\perp\!\!\!\perp Y$ but independent conditioning on Z, $X \perp\!\!\!\perp Y|Z$ so that $I(X;Y|Z) = 0$.

The second simulation (Figure 4.3) is from [66]: $X \sim$ Discrete Uniform$(0, 3)$ and $Y \sim$ Continuous Uniform$(X, X + 2)$ with an additional, independently generated random variable, $Z \sim$ Binomial$(3, 0.5)$. Here, $X \not\perp\!\!\!\perp Y|Z$ and $I(X;Y|Z) = \log 3 - 2\log 2/2$. This example, a combination of discrete and continuous random variables, is common in many applications. Here, the discrete variables are numeric but it is also reasonable to use the discrete distance metric for non-numeric categorical variables.

The third simulation (Figure 4.4) places probability mass of 0.4 at both $(1, 1)$, and $(-1, -1)$ and probability mass of 0.1 at $(1, -1)$ and $(-1, 1)$ with an independently generated $Z \sim$ Poisson(2). In this case, $X \not\perp\!\!\!\perp Y|Z$ with $I(X;Y|Z) = 2 \cdot 0.4 \log(0.4/0.5^2) + 2 \cdot 0.1 \log(0.1/0.5^2)$. In this example, all variables are discrete.

The fourth simulation is also from [66]. X and Y are a mixture distribution where with probability $\frac{1}{2}$, (X, Y) is multivariate Gaussian with a correlation

coefficient of 0.8 and with probability $\frac{1}{2}$, (X, Y) places probability mass of 0.4 at $(1, 1)$, and $(-1, -1)$ and probability mass of 0.1 at $(1, -1)$ and $(-1, 1)$, as in the third experiment. Z is an independently generated Binomial$(3, 0.2)$ so that $I(X; Y|Z) = I(X; Y)$. We separate the domain of the integral into its discrete and continuous parts; that is, $(1, 1)$, $(-1, -1)$, $(1, -1)$ and $(-1, 1)$ make up the discrete part and everywhere else the continuous part. From here we calculate MI on each partition by multiplying the distribution by $\frac{1}{2}$, yielding $I(X; Y|Z) = 0.4 \log(2 \cdot 0.4/0.5^2) + 0.1 \log(2 \cdot 0.1/0.5^2) + 0.125 \log(4/(1 - 0.8^2))$. Results are in Figure 4.5. In HIV research, for example HIV viral load, the amount of virus in a milliliter of a patient's blood can only be measured to a minimum threshold. Below that threshold, depending on the assay used, a patient is said to be undetectable. This is a real-world example of a random variable that is itself a mix of discrete and continuous, difficult for most regression models. The this experiment shows that the proposed estimator has no problem in this scenario.

The following figure, figure 4.6, shows performance on real data, and perhaps an abbreviated example for a use of information theoretic statistics on applications that typically use regression. Journalists at ProPublica curated and analyzed these data primarily to expose racial bias within the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm used "to assess a criminal defendant's likelihood of becoming a recidivist–a term used to describe criminals who re-offend" [69]. The data contain more than 10,000 records from individual criminal defendants in Broward County, Florida within 2013 and 2014. We focus on three variables, the defendant's race, record of recidivism—defined as a new arrest within two years, and the COMPAS decile (1–10 integer) score, which is intended to predict a defendant's likelihood of recidivism, a greater score indicates a greater risk of recidivism. More details on the data can be found at [69]. We excluded observations with no decile score, or if the time between the score and arrest date differed by more than 90 days, leaving 10010 records; this screen is more inclusive

Table 4.1: ProPublica COMPAS/Recidivism data summary by race

Race	Count	COMPAS Score (mean)	Recidivism (%)
African-American	4960	5.25	40.6
Asian	50	2.72	20.0
Caucasian	3503	3.59	28.8
Hispanic	885	3.25	25.4
Native American	31	4.51	32.3
Other	581	2.77	24.8

than that originally used. Table 4.1 shows an aggregated data summary by race.

We examine the relationship between race and recidivism controlling for the COMPAS decile score. It is easy to see from table 4.1 that race is associated with recidivism. Interestingly, the COMPAS score, which is computed using a questionnaire, does not explicitly ask for race. CMI, in this setting, quantifies the additional information in race for predicting recidivism that is not already contained in the COMPAS score. Said differently, is COMPAS indirectly using race to predict recidivism?

Figure 4.6 was generated similarly to figures 4.2 through 4.5; rather than using simulated datasets, we used bootstrap samples of the required sample size from the larger dataset to visualize each estimator's empirical distribution. While the true CMI in this case is unknown, it is possible to glean meaning from the estimates. Because the FP estimator was not intended for this type of data, it is not likely to be accurate. As previously mentioned in § 4.3, both RAVK1 and 2 are likely to underestimate CMI in small samples, which is clearly happening in this case because since CMI is non-negative and the violin plots are entirely below zero, though the upward trend indicates improvement with larger samples. The proposed estimator is fairly constant, with the violin plots flanking zero consistently. Table 4.2 provides CMI estimates for each estimator on the full data of 10010 observations. Given the performance of the proposed estimator on simulated data, it seems likely that zero is the true CMI. Thus, any information captured in race for predicting recidivism

Table 4.2: Estimated $I(\text{Race}; \text{Recidivism}|\text{COMPAS Score})$ on all observations

Method	FP	RAVK1	RAVK2	Proposed
Estimated CMI	2.4320	-0.0124	-0.0096	-0.0015

is already captured by COMPAS. This indicates that while not directly asking for race, COMPAS is using race to make predictions.

4.5 Conclusion

We have presented a non-parametric estimator of CMI (and MI) for arbitrary combinations of discrete, continuous, and mixed variables. Under mild assumptions, the estimator is consistent, and on empirical simulations, the proposed estimator performs better over other similar estimators in all sample-sizes. Note that we do not provide a mean squared error (MSE) convergence rate for the proposed estimator. We believe that in order to obtain such a rate, it will be necessary to make stronger assumptions than those we have made for both the discrete and continuous components of the joint distribution. Specifically, the discrete points in a distribution will likely require upper and lower bounds on their corresponding probabilities as in [48] while the continuous part will require some smoothness and tail assumptions as in [60]. While in practice, the necessary assumptions for the underlying distribution of the data are rarely verifiable, the theoretical guarantees would be helpful to understand this estimator's behavior in a high-dimensional regime beyond what we provide in theorem 4.3.3.

An analytical estimation of the sampling distribution for this estimator would also be of particular interest and utility for applied disciplines. Though, this development will likely require even stronger assumptions on the underlying distribution. Approximation with the bootstrap may be a way forward, as in [70]. Further development on approximating the sampling distribution of $\widehat{I}$ is an interesting, valuable open problem for further inference including testing and confidence intervals.

Another area of interest is estimation on data with strongly dependent variables. To control the bias emerging from strong dependence, we assume that the density is bounded. However, bias is likely to increase as this bound increases as one can see in the proof of theorem 4.3.1. One way to diminish bias arising from strong dependence may be to include an appropriate adjustment term, as in [71]. The correction in[71] may improve accuracy in such a setting and remains an area of future research.

Finally, MI and CMI are typically used to determine a relationship between two variables, or groups of variables. Of some interest is understanding many relationships, usually causal, between multiple variables. Expanding mutual information to many variables, Carrara and Vanslette [72] showed that any correlation estimator with some constraints obtain an information theoretic form. The RAVK estimator from [67] and others such as in reference [73] provide a paradigm for using information-theoretic measures to explore more complex relationships. Estimators of this genre could be helpful for exploring causal pathways between many variables.

In this work, we have attempted to provide an estimator for CMI/MI and show its behavior with limited distributional assumptions. However, stronger assumptions and modifications to the estimator itself will likely yield helpful information for practice and are certainly welcome. While there is room for development on several fronts, we believe that the estimator proposed here will be of current value to many applied researchers for quantifying marginal and conditional dependence between variables. The development of this estimator was primarily motivated by data from scientific applications. There are clear advantages of using CMI for scientific inquiry that we reiterate. While this method does require independent, identically distributed data, it does not require parametric assumptions or specific functional relationships between variables such as linearity to quantify dependence. Due to the data processing inequality, greater shared information among random variables indicates closer causal proximity in causal chains. In this vein, CMI (or MI) estimates

close to or equal to zero indicate likely conditional (or marginal) independence. For these reasons, information is ideal for inference and discovery causal of relationships which may not satisfy parametric requirements. And, like regression, information is easily interpretable: CMI, $I(X;Y|Z)$, can be understood as the degree of association or statistical dependence shared between X and Y given Z or controlling for Z. Finally, we encourage others to continue researching innovative methodologies to accommodate fields whose data is too messy for most current data science methodologies.

4.6 Appendix

4.6.1 Proof of Theorem 4.3.1

Proof. Define $f \equiv \frac{dP_{XY|Z}}{d\left(P_{X|Z} \times P_{Y|Z}\right)}$ and for a random variable W on $\mathcal{W}$ with probability measure, P_W, and $w \in \mathcal{W}$, set

$$P_W(w,r) = P_W\left(\{v \in \mathcal{W} : \|v-w\|_\infty \leq r\}\right). \tag{4.26}$$

Let $(x_1,y_1,z_1),\ldots,(x_n,y_n,z_n)$ be an i.i.d. random sample from P_{XYZ} and that $\widehat{I}_n(X,Y|Z)$ is the value of $\widehat{I}_{\text{prop}}(X,Y|Z)$ for this sample.

Partition $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ into three disjoint sets:

1. $\Omega_1 = \{(x,y,z) : f = 0\}$
2. $\Omega_2 = \{(x,y,z) : f > 0, P_{XYZ}(x,y,z,0) > 0\}$
3. $\Omega_3 = \{(x,y,z) : f > 0, P_{XYZ}(x,y,z,0) = 0\}$

so that $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z} = \Omega_1 \cup \Omega_2 \cup \Omega_3$. Notice that

$$\mathbb{E}\left[\widehat{I}_n(X,Y|Z)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\right] = \mathbb{E}\left[\xi_1\right] \tag{4.27}$$

so that we only need show that $\mathbb{E}[\xi_i] \to I(X;Y|Z)$ for one point. In light of this, we drop the subscript. Using the law of total expectation and properties of integrals,

$$
\begin{aligned}
&|\mathbb{E}[\xi] - I(X;Y|Z)| && (4.28)\\
&= \left| \mathbb{E}_{XYZ}\left[\mathbb{E}[\xi|X,Y,Z]\right] - \int f(x,y,z) dP_{XYZ}(x,y,z) \right| && (4.29)\\
&\leq \int |\mathbb{E}[\xi|x,y,z] - f(x,y,z)| \, dP_{XYZ}(x,y,z) && (4.30)\\
&= \int_{\Omega_1} |\mathbb{E}[\xi|x,y,z] - f(x,y,z)| \, dP_{XYZ}(x,y,z) && (4.31)\\
&\quad + \int_{\Omega_2} |\mathbb{E}[\xi|x,y,z] - f(x,y,z)| \, dP_{XYZ}(x,y,z) && (4.32)\\
&\quad + \int_{\Omega_3} |\mathbb{E}[\xi|x,y,z] - f(x,y,z)| \, dP_{XYZ}(x,y,z). && (4.33)
\end{aligned}
$$

For clarification, the value of $\mathbb{E}[\xi|X,Y,Z]$ depends on both the value of the value of the random vector (X,Y,Z) and rest of the sample. We show that

$$
\int_{\Omega_i} |\mathbb{E}[\xi|x,y,z] - f(x,y,z)| \, dP_{XYZ} \to 0 \tag{4.34}
$$

for each $i = 1,2,3$ in three cases.

Case1: $(x,y,z) \in \Omega_1$. Let $\pi_{XY}(\Omega_1) = \{(x,y) : (x,y,z) \in \Omega_1\}$ be the projection onto the first two coordinates of Ω_1. Using the definition of f as the RN derivative,

$$
\begin{aligned}
&P_{XY|Z}(\pi_{XY}(\Omega_1))\\
&= \int_{\pi_{XY}(\Omega_1)} f d(P_{X|Z} \times P_{Y|Z})\\
&= \int_{\pi_{XY}(\Omega_1)} 0 d(P_{X|Z} \times P_{Y|Z}) = 0.
\end{aligned}
$$

Then $P_{XYZ}(\Omega_1) = (P_{XY|Z} \times P_Z)(\Omega_1) = 0$. So,

$$
\int_{\Omega_i} |\mathbb{E}[\xi|x,y,z] - f(x,y,z)| \, dP_{XYZ} = 0. \tag{4.35}
$$

Case 2: Assume $(x, y, z) \in \Omega_2$. This is the partition of discrete points because singleton have positive measure in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Using lemma 4.6.8, we have

$$f(x, y, z) = \frac{P_{XYZ}(x, y, z, 0)P_Z(x, y, z, 0)}{P_{XZ}(x, y, z, 0)P_{YZ}(x, y, z, 0)}. \tag{4.36}$$

Knowing the exact value of f allows us to work with it directly.

Let ρ be the distance from (x, y, z) to its kNN. Proceed in two cases, when $\rho = 0$ and when $\rho > 0$ by writing the integrand as dominated by the following two terms:

$$\begin{aligned}
&|\mathbb{E}[\xi|x, y, z] - \log f(x, y, z)| \\
&\leq |\mathbb{E}[\xi|x, y, z, \rho > 0] - \log f(x, y, z)|\, \mathbb{P}(\rho > 0) \\
&\quad + |\mathbb{E}[\xi|x, y, z, \rho = 0] - \log f(x, y, z)|\, \mathbb{P}(\rho = 0) \\
&\equiv |\mathbb{E}[\xi|\rho > 0] - \log f|\, \mathbb{P}(\rho > 0) \\
&\quad + |\mathbb{E}[\xi|\rho = 0] - \log f|\, \mathbb{P}(\rho = 0)
\end{aligned}$$

suppressing the x, y, z for brevity. We bound $|\mathbb{E}[\xi|\rho > 0] - \log f|$ and $\mathbb{P}(\rho = 0)$ and show that $|\mathbb{E}[\xi|\rho = 0] - \log f|$ and $\mathbb{P}(\rho > 0)$ converge to zero. By proposition 4.6.1, there exist a finite set of points with positive measure $E \subseteq \Omega_2$ such that

$$P_{XYZ}(\Omega_2 \backslash E) < \frac{\epsilon}{3(4 \log n + \log C)}. \tag{4.37}$$

Starting with $\mathbb{P}(\rho > 0)$, $\rho > 0$ when less than k points in the sample equal (x, y, z). The number of points exactly equal to (x, y, z) has a binomial distribution with parameters, $n - 1$ and $P_{XYZ}(x, y, z, 0) \equiv P_{XYZ}(0)$, Binomial$(n - 1, P_{XYZ}(0))$. Because $\frac{k}{n} \to 0$ as $n \to \infty$, there must be an n sufficiently large such that

$$\begin{aligned}
&\max\left\{\frac{k}{n}, \frac{-2}{n} \log\left(\frac{\epsilon}{3(4 \log n + \log C)|E|}\right) + \frac{2k}{n}\right\} \\
&\leq \min_{(x,y,z)\in E} P_{XYZ}(x, y, z, 0).
\end{aligned}$$

This inequality ensures that $k-1 \leq (n-1)P_{XYZ}(x,y,z,0)$ for all $(x,y,z) \in E$ to use Chernoff's inequality [74, §2.2]:

$$\begin{aligned}
\mathbb{P}(\rho > 0) &= \mathbb{P}(\text{Binomial}(n-1, P_{XYZ}(0)) \leq k-1) \\
&\leq \exp\left\{\frac{-[(n-1)P_{XYZ}(0) - (k-1)]^2}{2P_{XYZ}(x,y,z,0)(n-1)}\right\} \\
&\leq \exp\left\{-\left(\frac{1}{2}nP_{XYZ}(0) - k\right)\right\} \\
&\leq \frac{\epsilon}{3(4\log n + \log C)|E|}.
\end{aligned}$$

To bound $|\mathbb{E}[\xi|\rho > 0] - \log f|$, first notice that $\tilde{k}, n_{XZ}, n_{YZ}, n_Z \leq n$. If $k = \tilde{k}$, then ξ uses ψ and if If $k < \tilde{k}$, then ξ uses log, so that $|\xi| \leq \max\{4\psi(n), 4\log n\} = 4\log n$. And, $f \leq C$ by assumption so $|\mathbb{E}[\xi|\rho > 0] - \log f| < 4\log n + \log C$.

Now we show that $|\mathbb{E}[\xi|\rho = 0] - \log f| \to 0$. When $\rho = 0$, there must be k or more points exactly equal to (x,y,z). Because a point in the sample being equal to (x,y,z) is an independent, Bernoulli event, and because when $\rho = 0$, $\tilde{k}$, defined in (4.19), will be the total number of points equal to (x,y,z), $\tilde{k} - k \sim \text{Binomial}(n - k - 1, P_{XYZ}(0))$. We can make identical arguments for $n_{XZ} - k, n_{YZ} - k$, and $n_Z - k$ in their respective subspaces so that $n_{XZ} - k \sim \text{Binomial}(n-k-1, P_{XZ}(0))$, $n_{YZ} - k \sim \text{Binomial}(n-k-1, P_{YZ}(0))$, and $n_Z - k \sim \text{Binomial}(n-k-1, P_Z(0))$. [66, Lemma B.2] provides a rigorous proof for this.

Showing that $|\mathbb{E}[\xi|\rho = 0] - \log f| \to 0$, we can choose k and n sufficiently large so that $\frac{1}{k} \leq \frac{\epsilon}{48|E|}$, $k \geq \frac{P_Z(0)}{1-P_Z(0)}$ and $\frac{k}{n} \leq \frac{\epsilon}{24|E|}$. Assume $\tilde{k} = k$, so that ξ will use ψ.

Using lemma 4.6.4 four times and that $P_{XYZ}(0) \leq P_{XZ}(0), P_{YZ}(0) \leq P_Z(0)$,

$$
\begin{aligned}
&|\mathbb{E}[\xi|\rho=0] - \log f| \\
&= \Big|\mathbb{E}[\psi(\tilde{k})|\rho=0] - \mathbb{E}[\psi(n_{XZ})|\rho=0] \\
&\quad - \mathbb{E}[\psi(n_{YZ})|\rho=0] + \mathbb{E}[\psi(n_Z)|\rho=0] \\
&\quad - \log\frac{(nP_{XYZ}(0))(nP_Z(0))}{(nP_{XZ}(0))(nP_{YZ}(0))}\Big| \\
&\leq \Big|\mathbb{E}[\psi(\tilde{k})|\rho=0] - \log nP_{XYZ}(0)\Big| \\
&\quad + |\mathbb{E}[\psi(n_{XZ})|\rho=0] - \log nP_{XZ}(0)| \\
&\quad + |\mathbb{E}[\psi(n_{YZ})|\rho=0] - \log nP_{YZ}(0)| \\
&\quad + |\mathbb{E}[\psi(n_Z)|\rho=0] - \log nP_Z(0)| \\
&\leq \frac{2}{k} + \frac{k}{nP_{XYZ}(0)} + \frac{2}{k} + \frac{k}{nP_{XZ}(0)} \\
&\quad + \frac{2}{k} + \frac{k}{nP_{YZ}(0)} + \frac{2}{k} + \frac{k}{nP_Z(0)} \\
&\leq \frac{8}{k} + \frac{4k}{nP_{XYZ}(0)} \\
&\leq \frac{\epsilon}{6|E|} + \frac{\epsilon}{6|E|P_{XYZ}(0)}.
\end{aligned}
$$

If $\tilde{k} > k$, ξ will use log and rather than ψ. Lemma 4.6.4 shows that the bound used above will also work in this case.

It is clear that $\mathbb{P}(\rho = 0) \leq 1$.

Putting together the previous parts,

$$
\begin{aligned}
&\int_{\Omega_2} |\mathbb{E}[\xi|x,y,z] - \log f(x,y,z)|\, dP_{XYZ}(x,y,z) \\
&= \sum_{(x,y,z)\in\Omega_2} |\mathbb{E}[\xi|x,y,z] - \log f(x,y,z)|\, P_{XYZ}(x,y,z,0) \\
&\equiv \sum_{(x,y,z)\in\Omega_2} |\mathbb{E}[\xi] - \log f|\, P_{XYZ}(0) \\
&= \sum_{(x,y,z)\in E} |\mathbb{E}[\xi] - \log f|\, P_{XYZ}(0) \\
&\quad + \sum_{(x,y,z)\in\Omega_2\setminus E} |\mathbb{E}[\xi] - \log f|\, P_{XYZ}(0) \\
&\leq \sum_{(x,y,z)\in E} |\mathbb{E}[\xi|\rho>0] - \log f|\, \mathbb{P}(\rho>0) P_{XYZ}(0) \\
&\quad + \sum_{(x,y,z)\in E} |\mathbb{E}[\xi|\rho=0] - \log f|\, \mathbb{P}(\rho=0) P_{XYZ}(0) \\
&\quad + \sum_{(x,y,z)\in\Omega_2\setminus E} |\mathbb{E}[\xi] - \log f|\, P_{XYZ}(0) \\
&\leq \sum_{(x,y,z)\in E} \log(n^4 C) \left(\frac{\epsilon}{3\log(n^4C)|E|}\right) P_{XYZ}(0) \\
&\quad + \sum_{(x,y,z)\in E} \left(\frac{\epsilon}{6|E|} + \frac{\epsilon}{6|E|P_{XYZ}(0)}\right) P_{XYZ}(0) \\
&\quad + \sum_{(x,y,z)\in\Omega_2\setminus E} (4\log n + \log C)\, P_{XYZ}(0) \\
&\leq \log(n^4C)|E| \left(\frac{\epsilon}{3(\log(n^4C))|E|}\right) + |E| \left(\frac{\epsilon}{3|E|}\right) \\
&\quad + P_{XYZ}\left(\Omega_2\setminus E\right) (4\log n + \log C) \\
&= \frac{\epsilon}{3} + \frac{\epsilon}{3} + \left(\frac{\epsilon}{3\log(n^4C)}\right) \log(n^4C) \\
&= \epsilon.
\end{aligned}
$$

Case 3: Assume $(x,y,z) \in \Omega_3$. This is the continuous partition because singletons have zero measure in $\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}$. Lemma 4.6.5 assures that $\tilde{k} \to k$ almost

surely as $n \to \infty$; $P_{XYZ}\left(\left\{(x,y,z) \in \Omega_3 : \tilde{k} \to k\right\}\right) = 1$. $\tilde{k}$ is discrete so there is an N such that for $n \geq N$, $\tilde{k} = k$ with probability one.

Define $F_\rho(r)$ as the cumulative distribution function of the kNN distance, r; that is, $F_\rho(r)$ is the probability that that the kNN distance is r or less. Begin by decomposing the integrand into its parts:

$$
\begin{aligned}
&|\mathbb{E}\left[\xi|X,Y,Z\right] - \log f(X,Y,Z)| && (4.38)\\
&\equiv |\mathbb{E}\left[\xi\right] - \log f| \\
&= \left| \int_0^\infty \left(\mathbb{E}\left[\xi|\rho = r\right] - \log f\right) dF_\rho(r) \right| \\
&= \Big| \int_0^\infty \mathbb{E}\left[\xi|\rho = r\right] - \log\left(\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)}\right) \\
&\quad + \log\left(\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)}\right) - \log f dF_\rho(r) \Big| \\
&= \Big| \int_0^\infty \mathbb{E}\left[\psi(k) - \psi(n_{XZ}) - \psi(n_{YZ}) - \psi(n_Z)|\rho = r\right] \\
&\quad - \log\left(\frac{(nP_{XYZ}(r))(nP_Z(r))}{(nP_{XZ}(r))(nP_{YZ}(r))}\right) \\
&\quad + \log\left(\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)}\right) - \log f dF_\rho(r) \Big| \\
&\leq \left| \int_0^\infty \psi(k) - \log(nP_{XYZ}(r)) dF_\rho(r) \right| && (4.39)\\
&\quad + \left| \int_0^\infty \mathbb{E}[\psi(n_{XZ})|\rho = r] - \log(nP_{XZ}(r)) dF_\rho(r) \right| && (4.40)\\
&\quad + \left| \int_0^\infty \mathbb{E}[\psi(n_{YZ})|\rho = r] - \log(nP_{YZ}(r)) dF_\rho(r) \right| && (4.41)\\
&\quad + \left| \int_0^\infty \mathbb{E}[\psi(n_Z)|\rho = r] - \log(nP_Z(r)) dF_\rho(r) \right| && (4.42)\\
&\quad + \left| \int_0^\infty \log\left(\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)}\right) - \log f dF_\rho(r) \right|. && (4.43)
\end{aligned}
$$

Next, we show that with sufficiently large n, each of these terms is less than $\epsilon/5$. Do to this, we change variables for each integral using lemma 4.6.6.

Beginning with (4.39),

$$\begin{aligned}
&\left|\int_0^\infty \psi(k) - \log(nP_{XYZ})(r)dF_\rho(r)\right| \\
&= \left|\psi(k) - \log n - \int_0^\infty \log P_{XYZ}(r)dF_\rho(r)\right| \\
&= \left|\psi(k) - \log n - \int_0^\infty P_{XYZ}(r)\frac{(n-1)!}{(k-1)!(n-k-1)!}\right. \\
&\quad \left.[P_{XYZ}(r)]^{k-1}[1-P_{XYZ}(r)]^{n-k-1}dP_{XYZ}(r)\right| \\
&= \left|\psi(k) - \log n - \frac{(n-1)!}{(k-1)!(n-k-1)!}\right. \\
&\quad \left.\int_0^\infty [P_{XYZ}(r)]^{k}[1-P_{XYZ}(r)]^{n-k-1}dP_{XYZ}(r)\right| \\
&= |\psi(k) - \log n - (\psi(k) - \psi(n))| \\
&= |\psi(n) - \log(n)| < \frac{1}{n}.
\end{aligned}$$

For lines 4.40, 4.41, and 4.42, consider the random variables n_{XZ}, n_{YZ}, and n_Z defined in line 4.15. In this case, we know that $\tilde{k} = k$ almost surely. Note that $n_{XZ}, n_{YZ}, n_Z \geq k$. Observation j in the sample will contribute to the count of $n_{W,i} - k$ for $W \in \{(XZ), (YZ), Z\}$ when $\|w_i - w_j\|_\infty \leq \rho_{k,i}$ given that it is not one of the first k nearest neighbors. There are $n - k - 1$ independent, identically distributed, data points left not counting the k nearest neighbors or point i. A point j has probability, $P_W(\rho_{k,i})$, that it is within a radius of $\rho_{k,i}$ in the W subspace. The probability that a point falls within a radius of $\rho_{k,i}$ in the XYZ-space is $P_{XYZ}(\rho_{k,i})$. Using basic conditional probability rules, one can see that the probability that any point contributes to the count of n_W is $\frac{P_W(\rho_{k,i}) - P_{XYZ}(\rho_{k,i})}{1 - P_{XYZ}(\rho_{k,i})}$. Then, for $W \in \{XZ, YZ, Z\}$

$$\begin{aligned}
&n_{W,i} - k \sim \\
&\quad \text{Binomial}\left(n - k - 1, \frac{P_W(\rho_{k,i}) - P_{XYZ}(\rho_{k,i})}{1 - P_{XYZ}(\rho_{k,i})}\right)
\end{aligned} \tag{4.44}$$

$P_{XYZ}(\rho_{k,i}) \leq P_W(\rho_{k,i})$ for all points. Choosing k such that $k \geq \frac{15+3\epsilon}{\epsilon}$ and applying lemma 4.6.9, we bound lines 4.40, 4.41, and 4.42 by $\frac{\epsilon}{5}$.

Moving to line 4.43, using lemma 4.6.7, we have

$$\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} \to f \tag{4.45}$$

(converges pointwise) as $r \to 0$ and

$$\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} \leq C \tag{4.46}$$

almost everywhere $[P_{X|Z} \times P_{Y|Z}]$. Using Egoroff's theorem, there exists a measurable set, $E \subseteq \Omega_3$ such that

$$P_{XYZ}(\Omega_3 \backslash E) \leq \frac{\epsilon}{10 \log C} \tag{4.47}$$

and

$$\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} \xrightarrow{U} f \tag{4.48}$$

(converges uniformly) as $r \to 0$ on E. Using the uniform convergence on E, there exists $r_\epsilon > 0$ such that for all $r \leq r_\epsilon$

$$\left| \log \frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f \right| \leq \frac{\epsilon}{20} \tag{4.49}$$

for all $(x, y, z) \in E$. And for sufficiently large n, we have

$$\max \left\{ \frac{k}{n}, \frac{-2 \log \left(\frac{\epsilon}{40 \log C} \right) + 2k}{n} \right\} \leq P_{XYZ}(r_\epsilon). \tag{4.50}$$

Consider the probability, $\mathbb{P}(\rho > r_\epsilon)$, that a point's kNN distance is greater than r_ϵ. This can only happen when $k-1$ or less neighbors fall within a radius of r_ϵ. There are $n-1$ independent, identically distributed points that can potentially fall in to this region each with probability, $P_{XYZ}(r_\epsilon)$ so that this also has a binomial

distribution. Again using Chernoff's inequality,

$$\begin{aligned}
&\mathbb{P}(\rho > r_\epsilon) \\
&\leq \exp\left(\frac{-[(n-1)P_{XYZ}(r_\epsilon) - (k-1)]^2}{2P_{XYZ}(r_\epsilon)(n-1)}\right) \\
&\leq \exp\left(-\frac{1}{2}nP_{XYZ}(r_\epsilon) + k\right) \\
&\leq \frac{\epsilon}{40\log C}.
\end{aligned}$$

With assumption 2, $f \leq C$, and line 4.46 from proposition 4.6.7,

$$\left|\log\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f\right| \leq 2\log C. \tag{4.51}$$

For points $(x, y, z) \in E$,

$$\begin{aligned}
&\left|\int_0^\infty \log\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f dF_\rho(r)\right| \\
&\leq \int_0^\infty \left|\log\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f\right| dF_\rho(r) \\
&= \int_0^{r_\epsilon} \left|\log\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f\right| dF_\rho(r) \\
&\quad + \int_{r_\epsilon}^\infty \left|\log\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f\right| dF_\rho(r) \\
&\leq \int_0^{r_\epsilon} \frac{\epsilon}{20} dF_\rho(r) + \int_{r_\epsilon}^\infty 2\log C dF_\rho(r) \\
&= \frac{\epsilon}{20}\mathbb{P}(\rho \leq r_\epsilon) + (2\log C)\mathbb{P}(\rho > r_\epsilon) \\
&\leq \frac{\epsilon}{20} + \frac{\epsilon}{20} = \frac{\epsilon}{10}.
\end{aligned}$$

But, for points $(x, y, z) \in \Omega_3 \backslash E$, it is only necessary bound the integrand,

$$\begin{aligned}
&\left| \int_0^\infty \log \frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f dF_\rho(r) \right| \\
&\leq \int_0^\infty \left| \log \frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f \right| dF_\rho(r) \\
&\leq \int_0^\infty 2 \log C dF_\rho(r) \\
&\leq 2 \log C.
\end{aligned}$$

The last step follows because $F_\rho(r)$ is a probability measure. Integrating term 4.43 over all of Ω_3,

$$\begin{aligned}
&\int_{\Omega_3} \left| \int_0^\infty \log \frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f dF_\rho(r) \right| dP_{XYZ} \\
&\leq \int_E \left| \int_0^\infty \log \frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f dF_\rho(r) \right| dP_{XYZ} \\
&+ \int_{\Omega_3 \backslash E} \left| \int_0^\infty \log \frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f dF_\rho(r) \right| dP_{XYZ} \\
&\leq \int_E \frac{\epsilon}{10} dP_{XYZ} + \int_{\Omega_3 \backslash E} 2 \log C dP_{XYZ} \\
&= \frac{\epsilon}{10} + (2 \log C) P_{XYZ}(\Omega \backslash E) \leq \frac{\epsilon}{5}
\end{aligned}$$

where we used Ergoroff's theorem from line 4.47 in the last line. Now we integrate line 4.38 over Ω_3 using the previous arguments showing that lines 4.39 4.40, 4.41, 4.42, and 4.43 are all bounded. Choosing n large enough to satisfy the previous conditions, we have

$$
\begin{aligned}
&\int_{\Omega_3} |\mathbb{E}\left[\xi\right] - \log f| \, dP_{XYZ} \\
&\leq \int_{\Omega_3} \left| \int_0^\infty \psi(k) - \log(nP_{XYZ}(r)) dF_\rho \right| dP_{XYZ} \\
&+ \int_{\Omega_3} \left| \int_0^\infty \mathbb{E}[\psi(n_{XZ})] - \log(nP_{XZ}(r)) dF_\rho \right| dP_{XYZ} \\
&+ \int_{\Omega_3} \left| \int_0^\infty \mathbb{E}[\psi(n_{YZ})] - \log(nP_{YZ}(r)) dF_\rho \right| dP_{XYZ} \\
&+ \int_{\Omega_3} \left| \int_0^\infty \mathbb{E}[\psi(n_Z)] - \log(nP_Z(r)) dF_\rho \right| dP_{XYZ} \\
&+ \int_{\Omega_3} \left| \int_0^\infty \log \frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} - \log f dF_\rho \right| dP_{XYZ} \\
&\leq \int_{\Omega_3} \frac{\epsilon}{5} dP_{XYZ} + \int_{\Omega_3} \frac{\epsilon}{5} dP_{XYZ} \\
&\quad + \int_{\Omega_3} \frac{\epsilon}{5} dP_{XYZ} + \int_{\Omega_3} \frac{\epsilon}{5} dP_{XYZ} + \frac{\epsilon}{5} \\
&= \epsilon
\end{aligned}
$$

□

4.6.2 Proof of Theorem 4.3.2

Proof. Let $W_1' \ldots, W_n'$ be another random sample of size n such that for each i, $W_i = (X_i, Y_i, Z_i)$, $W_i' = (X_i', Y_i', Z_i')$ and that $W_i \stackrel{d}{=} W_i'$ (equally distributed).
Let $W^{(i)} = \{W_1, \ldots, W_{i-1}, W_i', W_{i+1}, \ldots, W_n\}$ and let $W^{i-} = \{W_1, \ldots, W_{i-1}, W_{i+1}, \ldots, W_n\}$.
We proceed using the Stein-Efron inequality as in [74, Theorem 3.1],

$$
\operatorname{Var}\left(\widehat{I}_n(W)\right) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[\widehat{I}_n(W) - \widehat{I}_n(W^{(i)})\right]^2.
$$

To reduce the number of cases we must examine, consider the following supre-

mum over possible values $w_1, \dots w_n, w_i'$ of the random vector W:

$$\begin{aligned}
&\sup_{w_1,\dots w_n,w_i'} \left|\widehat{I}_n(W) - \widehat{I}_n(W^{(i)})\right| \\
&\leq \sup_{w_1,\dots w_n,w_i'} \Big(\left|\widehat{I}_n(W) - \widehat{I}_n(W^{i-})\right| \\
&\quad + \left|\widehat{I}_n(W^{i-}) - \widehat{I}_n(W^{(i)})\right|\Big) \\
&\leq \sup_{w_1,\dots w_n} \left|\widehat{I}_n(W) - \widehat{I}_n(W^{i-})\right| \\
&\quad + \sup_{w_1,\dots,w_{i-1},w_i',w_{i+1},\dots,w_n} \left|\widehat{I}_n(W^{i-}) - \widehat{I}_n(W^{(i)})\right| \\
&= 2 \sup_{w_1,\dots w_n} \left|\widehat{I}_n(W) - \widehat{I}_n(W^{i-})\right| \\
&= \frac{2}{n} \sup_{w_1,\dots w_n} \sum_{j=1}^{n} \left|\xi_j(W) - \xi_j(W^{i-})\right|.
\end{aligned}$$

The penultimate step holds because $W \stackrel{d}{=} W^{(i)}$.

We proceed by bounding $\left|\xi_j(W) - \xi_j(W^{i-})\right|$ by looking at the individual cases.

Case 1: $i = j$.

Notice that if $0 < a, b \leq n$ then

$$\begin{aligned}
&|\psi(a) - \log(b)| \leq |\psi(a) - \log(a)| + |\log(b) - \log(b)| \\
&\leq \frac{1}{b} + \log(\max\{a, b\}) \leq \log n + 1.
\end{aligned}$$

Using this,

$$\begin{aligned}
&\left|\xi_j(W) - \xi_j(W^{i-})\right| \\
&\leq \left|\psi(k) - \log(\tilde{k}'_j)\right| + \left|\psi(n_{XZ,j}) - \log(n'_{XZ,j})\right| \\
&\quad + \left|\psi(n_{YZ,j}) - \log(n'_{YZ,j})\right| + \left|\psi(n_{Z,j}) - \log(n'_{Z,j})\right| \\
&\leq 4\log n + 4.
\end{aligned}$$

In the summation from $j = 1$ to n, this can only happen one times, so we have that $\sum_{j=1}^{n} \left|\xi_j(W) - \xi_j(W^{i-})\right| \leq 4\log n + 4$.

Case 2: $i \neq j, \tilde{k}_j > k$.

Recall that $\xi_j(W) = \log(\tilde{k}_j) - \log(n_{XZ,j}) - \log(n_{YZ,j}) + \log(n_{Z,j})$ and that $\rho_{k,j}$ is the ℓ_∞-distance from W_j to its kNN. Removing W_i from W will only change $\xi_j(W)$ if W_i is counted in $\tilde{k}_j, n_{XZ,j}, n_{YZ,j}$, or $n_{Z,j}$. Because $\tilde{k}_j > k$, there must be at least two points whose distance to W_j is exactly $\rho_{k,j}$, so removing one point cannot change $\rho_{k,j}$, regardless of its location with respect to W_j. Because $\rho_{k,j}$ will remain unchanged after removing W_i from W, $\tilde{k}_j, n_{XZ,j}, n_{YZ,j}$, or $n_{Z,j}$ can each only decrease by a count of one. Under $\xi_j(W^{i-})$, if $\tilde{k}_j = k$, then the *log* function will become ψ. In general, we have that $\psi(w) - \psi(w-1) = \frac{1}{w-1}$, $\log(w) - \log(w-1) = \log\left(\frac{w}{w-1}\right) \leq \frac{1}{w-1}$

and, $\log(w) - \psi(w-1) = \log(w) - \psi(w) + \frac{1}{w-1} \leq \frac{2}{w-1}$. Regardless, we have

$$\begin{aligned}
&\left|\xi_j(W) - \xi_j(W^{i-})\right| \\
&\leq \left|\log(\tilde{k}_j) - \psi(\tilde{k}_j - 1)\right| \\
&\quad + \left|\log(n_{XZ,j}) - \psi(n_{XZ,j} - 1)\right| \\
&\quad + \left|\log(n_{YZ,j}) - \psi(n_{YZ,j} - 1)\right| \\
&\quad + \left|\log(n_{Z,j}) - \psi(n_{Z,j} - 1)\right| \\
&\leq \frac{2}{\tilde{k}_j - 1} + \frac{2}{n_{XZ,j} - 1} \\
&\quad + \frac{2}{n_{YZ,j} - 1} + \frac{2}{n_{Z,j} - 1}.
\end{aligned}$$

Now, rather than considering the number of points that can change with the removal of W_i, we focus on the number of counts, $\tilde{k}_j, n_{XZ,j}, n_{YZ,j}$, and $n_{Z,j}$, that will change. If W_i is among the $\tilde{k}_j$NN of W_j, then its removal can change at most the $\tilde{k}_j$ points within a distance of $\rho_{k,j}$ in all coordinates. If W_i is not among the $\tilde{k}_j$NN of W_j but is counted in $n_{XZ,j}$, (and possibly in $n_{Z,j}$ too), then its removal will not affect $\tilde{k}_j$ or $n_{YZ,j}$ and will only change $n_{XZ,j}$, (and $n_{Z,j}$) for the points within a distance of $\rho_{k,j}$ from W_j in the XZ coordinates, which is $n_{XZ,j}$. Similarly, $n_{YZ,j}$

and $n_{Z,j}$ will change for at most $n_{YZ,j}$ and $n_{Z,j}$ points, respectively. So, we have

$$\begin{aligned}
&\sum_{j=1}^{n} \left|\xi_j(W) - \xi_j(W^{(i)})\right| \\
&\leq \sum_{j=1}^{n} \frac{2}{\tilde{k}_j - 1} + \sum_{j=1}^{n} \frac{2}{n_{XZ,j} - 1} \\
&\quad + \sum_{j=1}^{n} \frac{2}{n_{YZ,j} - 1} + \sum_{j=1}^{n} \frac{2}{n_{Z,j} - 1} \\
&\leq \frac{2\tilde{k}_j}{\tilde{k}_j - 1} + \frac{2n_{XZ,j}}{n_{XZ,j} - 1} \\
&\quad + \frac{2n_{YZ,j}}{n_{YZ,j} - 1} + \frac{2n_{Z,j}}{n_{Z,j} - 1} \\
&\leq 16
\end{aligned}$$

Case 3: $i \neq j, \tilde{k}_j = k$.

Again, removing W_i from W will change $\xi_j(W)$ only if W_i is counted in at least one of $\tilde{k}_j, n_{XZ,j}, n_{YZ,j}$, or $n_{Z,j}$. If W_i is within the kNN of W_j, then removing W_i will change the value of $\rho_{k,j}$. Because $\rho_{k,j}$ is different, we cannot say how $n_{XZ,j}, n_{YZ,j}$, or $n_{Z,j}$ will change so we give the loosest bound from case 1:

$$\left|\xi_j(W) - \xi_j(W^{i-})\right| \leq 4 \log n + 4.$$

Using the first part of [66, Lemma C.1], if $U_i', U_1, \ldots, U_n$ are vectors in $\mathbb{R}^d$ and $\mathbf{U} = \{U_1, \ldots, U_{j-1}, U_i', U_{j+1}, \ldots, U_n\}$, then

$$\sum_{j=1}^{n} I_{\left\{U_i' \text{ is in the } k\text{NN of } U_j \text{ in } \mathbf{U}\right\}} \leq k\gamma_d$$

where γ_d is a constant that only depends on the dimension of the XYZ space [75,

Corollary 6.1]. With this, we have

$$\sum_{i=1}^{n} \left[\xi_j(W) - \xi_j(W^{i-})\right] \leq k\gamma_d(4\log n + 4).$$

If W_i is not within the kNN of W_j, it can still contribute to the count of $n_{XZ,j}, n_{YZ,j}$, or $n_{Z,j}$. In this case $\rho_{k,j}$ will not change, so removing one point will decrease $n_{XZ,j}, n_{YZ,j}$, or $n_{Z,j}$ by at most one, similar to case 2.

$$\begin{aligned}
&\left|\xi_j(W) - \xi_j(W^{(i)})\right| \\
&\leq |\psi(k) - \psi(k)| \\
&\quad + |\psi(n_{XZ,j}) - \psi(n_{XZ,j} - 1)| \\
&\quad + |\psi(n_{YZ,j}) - \psi(n_{YZ,j} - 1)| \\
&\quad + |\psi(n_{Z,j}) - \psi(n_{Z,j} - 1)| \\
&= \frac{1}{n_{XZ,j} - 1} + \frac{1}{n_{YZ,j} - 1} \\
&\quad + \frac{1}{n_{Z,j} - 1}.
\end{aligned}$$

Using the second part of [66, Lemma C.1], if $U_i', U_1, \ldots, U_n$ are vectors in $\mathbb{R}^d$ and $\mathbf{U} = \{U_1, \ldots, U_{j-1}, U_i', U_{j+1}, \ldots, U_n\}$, then

$$\sum_{j=1}^{n} \frac{1}{k_i} I_{\left\{U_i' \text{ is in the } k_i\text{NN of } U_j \text{ in } \mathbf{U}\right\}} \leq \gamma_d(\log n + 1).$$

Then

$$\begin{aligned}
&\sum_{j=1}^{n} \left|\xi_j(W) - \xi_j(W^{i-})\right| \\
&\leq \sum_{j=1}^{n} \frac{1}{n_{XZ,j} - 1} + \sum_{j=1}^{n} \frac{1}{n_{YZ,j} - 1} \\
&\quad + \sum_{j=1}^{n} \frac{1}{n_{Z,j} - 1} \\
&\leq \sum_{j=1}^{n} \frac{1}{n_{XZ,j}} + \sum_{j=1}^{n} \frac{1}{n_{YZ,j}} \\
&\quad + \sum_{j=1}^{n} \frac{1}{n_{Z,j}} + 3 \\
&\leq (\gamma_{d_{XZ}})(\log n + 1) + \gamma_{d_{YZ}}(\log n + 1) \\
&\quad + \gamma_{d_Z}(\log n + 1) + 3 \\
&\leq \gamma_d(\log n + 1) + 3
\end{aligned}$$

where d_{XZ} is the dimension of XZ, etc.

Combining all of these cases, we have

$$\begin{aligned}
&\sum_{j=1}^{n} \left|\xi_j(W) - \xi_j(W^{(i)})\right| \\
&\leq (4 \log n + 4) + 16 + k\gamma_d(4 \log n + 4) \\
&\quad + \gamma_d(\log n + 1) + 3 \\
&\leq 36k\gamma_d \log n
\end{aligned}$$

for $n \geq 2, k \geq 1$ (and $d \geq 3$ so $\gamma_d \geq 3$). Using Stein-Efron inequality,

$$\begin{aligned}
&\mathrm{Var}\left(\widehat{I}_n(W)\right) \\
&\quad \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\widehat{I}_n(W) - \widehat{I}_n(W^{(i)})\right]^2 \\
&\quad = \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^{n}\xi_j(W) - \frac{1}{n}\sum_{j=1}^{n}\xi_j(W^{(i)})\right|^2 \\
&\quad \leq \frac{1}{2n^2}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{j=1}^{n}\left|\xi_j(W) - \xi_j(W^{(i)})\right|\right]^2 \\
&\quad \leq \frac{1}{2n^2}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{j=1}^{n}\sup_{W}\left|\xi_j(W) - \xi_j(W^{i-})\right|\right]^2 \\
&\quad \leq \frac{1}{2n^2}\sum_{i=1}^{n}\mathbb{E}\left[36k\gamma_d\log n\right]^2 \\
&\quad = \frac{648k^2\gamma_d^2(\log n)^2}{n} \\
&\quad \to 0.
\end{aligned}$$

The last step uses l'Hospital's rule twice. □

4.6.3 Proof of Corollary 2

Proof. From the proof in theorem 4.3.2, it is easy to verify that

$$\begin{aligned}
&\sup_{w_1,\ldots w_n,w_i'}\left|\widehat{I}_n(W) - \widehat{I}_n(W^{(i)})\right| \\
&\qquad \leq \frac{2}{n}\sup_{w_1,\ldots w_n}\sum_{j=1}^{n}\left|\xi_j(W) - \xi_j(W^{i-})\right| \\
&\qquad \leq \frac{72k\gamma_d\log n}{n}.
\end{aligned}$$

So, $\widehat{I}_n$ satisfies the bounded difference property. Using the bounded difference inequality ([74, Theorem 6.2]) with

$$v = \frac{1296k^2(\log n)^2}{n},$$

we bound the one-sided probability by $\exp\left\{-t^2/(2v)\right\}$ and simply multiply this value by a factor of 2. □

4.6.4 Proof of Theorem 4.3.3

Proof. Let (x, y, z) be an arbitrary point in the domain of (X, Y, Z). Choose $r \geq 0$ if (x, y, z) is a discrete point and $r > 0$ if (x, y, x) is a continuous point. Recall that we define $P_Z(r) \equiv P_Z(B(z, r))$. Proceeding by contradiction, assume that $\lim_{d\to\infty} P_Z(r) > 0$; that is, there exists a $\delta > 0$ such that for every $D > 0$, there is a $d \geq D$ such that $P_Z(r) > \delta$. $B(z, r)$ is a d-dimensional, ℓ_∞-ball so it can be written as the product of d sets. Defining $Z^k \equiv (Z_{k-1}, \cdots, Z_1)$ for $k = 1, 2, \ldots, d$. $P_{Z_k|Z^k}(r) \equiv \mathbb{P}(Z_k \in \pi_k(B(z, r))|Z^k \in \pi^k(B(z, r)))$ where π_k is the projection on to the kth coordinate and π^k is the projection on to the $k-1, \ldots, 1$ coordinates. Then we have that

$$\prod_{k=1}^{d} P_{Z_k|Z^k}(r) = P_Z(r) > \delta.$$

Then

$$\lim_{d\to\infty} \sum_{k=1}^{d} \log P_{Z_k|Z^k}(r) > \log \delta > -\infty.$$

For each k, $\log P_{Z_k|Z^k}(r) \leq 0$ so $\log P_{Z_k|Z^k}(r) \to 0$ as $d \to \infty$ using the fact that $a_i \geq 0, \sum_{i=1}^{\infty} a_i < M$ for some $M \Rightarrow a_i \to 0$.

Choose $\epsilon > 0$ and let $\mathcal{Q}$ be a finite partition of the domain of Z into sets with positive measure in P_Z. Because z and r were chosen arbitrarily in the previous part, then for each $Q \in \mathcal{Q}$, there is a point z_Q in the domain of Z and distance r_Q such that $B(z_Q, r_Q) \subseteq Q$. Then there must be a d_Q such that for every $k \geq$

d_Q, $-\log P_{Z_k|Z^K}(B(z_Q, r_Q)) \leq \frac{\epsilon}{\|\mathcal{Q}\|}$ because $\log P_{Z_k|Z^K}(B(z_Q, r_Q)) \to 0$ for each Q. Choosing $k \geq \max_{Q\in\mathcal{Q}} d_Q$, we have that

$$\begin{aligned}
&\sum_{Q\in\mathcal{Q}} -P_{Z_kZ^k}(Q)\log P_{Z_k|Z^k}(Q) \\
&\leq \sum_{Q\in\mathcal{Q}} -\log P_{Z_k|Z^k}(Q) \\
&\leq \sum_{Q\in\mathcal{Q}} -\log P_{Z_k|Z^K}(B(z_Q, r_Q)) \\
&\leq \sum_{Q\in\mathcal{Q}} \frac{\epsilon}{\|\mathcal{Q}\|} = \epsilon.
\end{aligned}$$

Let $\{\mathcal{Q}_l : l = 1, 2, \dots\}$ be a sequence of increasingly fine partitions of the domain of Z into sets with positive measure in P_Z. Using [45, Lemma 7.18], we have that

$$H(Z_k|Z^k) = \lim_{l\to\infty} \sum_{Q\in\mathcal{Q}_l} -P_{Z_kZ^k}(Q)\log P_{Z_k|Z^k}(Q) \leq \epsilon.$$

Using Cesàro's lemma ($a_i \to a \Rightarrow \frac{1}{n}\sum_{i=1}^n a_i \to a$),

$$\lim_{d\to\infty} \frac{1}{d} H(Z) = \lim_{d\to\infty} H(Z_d|Z^d) \leq \epsilon.$$

But, ϵ was chosen arbitrarily, so

$$\lim_{d\to\infty} \frac{1}{d} H(Z) = 0,$$

a contradiction. Thus, $\lim_{d\to\infty} P_Z(r) = 0$ for all z in the domain of Z.

Again, by contradiction, assume that $P_Z(\rho_k) \xrightarrow{P} 1$ as $d \to \infty$. Then

$$\begin{aligned}
\sum_{d=1}^{\infty} \log P_{Z_d|Z^d}(\rho_k) &= \log\left(\prod_{l=d}^{\infty} P_{Z_d|Z^d}(\rho_k)\right) \\
&= \log\left(P_Z(\rho_k)\right) \xrightarrow{P} 0
\end{aligned}$$

using the continuous mapping theorem. But, the sum of non-positive values can converge to zero only if $\log P_{Z_d|Z^d}(\rho_k) = 0$ for each d with probability one. Then $P_Z(\rho_k) = 1$ for each finite d.

Fix d. For $P_Z(\rho_k) = P_Z(B(z, \rho_k)) = 1$, $B(z, \rho_k)$ must include the support of Z. Then kNN (in the XYZ space) must be on a boundary of the domain of Z and ρ_k, the ℓ_∞, kNN distance in XYZ, must be at least half of diameter of the domain of Z with probability one. Because all observations are independent of each other and identically distributed, all Z-coordinates within the sample must also be on the boundary of the domain of Z with probability one. If Z were continuous, then the boundary would have measure zero, indicating that each coordinate of Z must be discrete. Note that Z coordinates need not be binary if using a discrete scalar distance metric for non-numeric, categorical variables. If the support of Z contains more than one point, then ties are possible with positive probability, and $\rho_k = 0$ with positive probability and $P_Z(B(z, \rho_k)) < 1$. Then Z must have support on one point, again contradicting a non-zero entropy rate for Z. This indicates that $\lim_{d\to\infty} P_Z(\rho_k) < 1$.

Using this fact, there must be an r such that for each $d \geq 1$, $P_Z(\rho_k) \leq P_Z(r) < 1$, so that $P_Z(\rho_k) \leq P_Z(r) \to 0$ as $d \to \infty$.

Finally, because $P_Z(\rho_k) \geq P_{XYZ}(\rho_k)$, we must have

$$\frac{P_Z(\rho_k) - P_{XYZ}(\rho_k)}{1 - P_{XYZ}(\rho_k)} \xrightarrow{P} 0$$

as $d \to \infty$. Recall that $n_Z - k$ has a binomial distribution with the probability parameter stated above which converges to zero. From here, it is easy to see that $n_Z \xrightarrow{D} k$ (converges in distribution) as $d \to \infty$. Because n_Z is converging to a constant, we also have $n_Z \xrightarrow{P} k$. But, $k \leq \tilde{k}, n_{XZ}, n_{YZ} \leq n_Z$, so $\tilde{k}, n_{XZ}, n_{YZ} \xrightarrow{P} k$ as well. By the continuous mapping theorem, for each sample point,

$$\xi_i = \psi(k) - \psi(n_{XZ}) - \psi(n_{YZ}) + \psi(n_Z) \xrightarrow{P} 0$$

so that

$$\widehat{I}_{\text{prop}}(X;Y|Z) = \frac{1}{n}\sum_{i=1}^{n}\xi_i \xrightarrow{P} 0.$$

□

4.6.5 Auxiliary Lemmas

Proposition 4.6.1. *Let $(X, 2^X, \mu)$ be a discrete measure space with $\mu(X) = C < \infty$. Then for every $\epsilon > 0$, there exists a finite set E such that $\mu(X\backslash E) < \epsilon$ and each point in E has non-zero measure.*

Proof. If X is finite, the problem is trivial. Assume X in infinite. Without loss of generality, remove any zero-measure points from X. Because $(X, 2^X, \mu)$ is discrete, X must be countable so we number each point in X. We must have that $\sum_{i=1}^{\infty}\mu(x_i) = C$. Then there must be a positive integer, N, such that for each $n \geq N$, $C - \sum_{i=1}^{n}\mu(x_i) < \epsilon$. Let $E = \{x_i : 1 \leq i \leq N\}$. Then $\mu(X\backslash E) = \mu(X) - \mu(E) = C - \sum_{i=1}^{N}\mu(x_i) < \epsilon$. □

Proposition 4.6.2. *Assume $W_n \sim$ Binomial(n, p), then*

$$\mathbb{E}\left[\frac{1}{W_n + 1}\right] = \frac{1 - (1-p)^{n+1}}{(n+1)p} \leq \frac{1}{np} \tag{4.52}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{W_n+1}\right] &= \sum_{m=0}^{n} \frac{1}{m+1}\binom{n}{m}p^m(1-p)^{m-n} \\
&= \frac{1}{(n+1)p}\sum_{m=0}^{n}\binom{n+1}{m+1}p^{m+1}(1-p)^{n-m} \\
&= \frac{1}{(n+1)p}\sum_{m=1}^{n}\binom{n+1}{m}p^{m}(1-p)^{n+1-m} \\
&= \frac{1}{(n+1)p}\left[1-\mathbb{P}(X_{n+1}=0)\right] \\
&= \frac{1-(1-p)^{n+1}}{(n+1)p}
\end{aligned}$$

□

Proposition 4.6.3. *Let $W \sim$ Binomial(n,p) then*

$$\left|\mathbb{E}\left[\log\left(\frac{W+k}{np+k}\right)\right]\right| \leq \frac{1}{np+k} \tag{4.53}$$

Proof. Using Taylor's theorem to expanding $\log(x)$ about $np+k$, there exists $c \in [x, np+k]$ such that

$$\log(x) = \log(np+k) + \frac{x-np-k}{np+k} - \frac{(x-np-k)^2}{2c^2}. \tag{4.54}$$

Plugging in $W+k$ for x and aggregating the log terms,

$$\log\left(\frac{W+k}{np+k}\right) = \frac{W-np}{np+k} - \frac{(W-np)^2}{2c^2}. \tag{4.55}$$

Taking the expected value of both sides, the first-order term drops out,

$$\mathbb{E}\left[\log\left(\frac{W+k}{np+k}\right)\right] = \mathbb{E}\left[-\frac{(W-np)^2}{2c^2}\right]$$

for some $c \in [np+k, W+k]$. Notice that $\mathbb{E}\left[\log\left(\frac{W+k}{np+k}\right)\right] \leq 0$ for all c, so that

$$\left|\mathbb{E}\left[\log\left(\frac{W+k}{np+k}\right)\right]\right| \leq \mathbb{E}\left[\max_{c\in[np+k,W+k]}\left\{\frac{(W-np)^2}{2c^2}\right\}\right]. \tag{4.56}$$

Because $\frac{1}{2c^2}$ is monotonic, $\frac{(W-np)^2}{2c^2}$ is optimized at the boundary values of $c = np+k$ and $c = W+k$. If $c = np+k$,

$$\begin{aligned}\mathbb{E}\left[\frac{(W-np)^2}{2c^2}\right] &= \frac{np(1-p)}{2(np+k)^2} \\ &\leq \frac{np+k}{2(np+k)^2} \\ &= \frac{1}{2(np+k)}\end{aligned}$$

using $\mathbb{E}[(W-np)^2] = \mathrm{Var}(W) = np(1-p)$.

If $c = W+k$ and $k \leq np$, we use $\sum_{j=0}^{n}\binom{n+2}{j+2}p^{j+2}(1-p)^{n-j} = \mathbb{P}(V \geq 2)$ where $V \sim \text{Binomial}(n+2, p)$, so that

$$\begin{aligned}&\mathbb{E}\left[\frac{(W-np)^2}{2(W+k)^2}\right] \\ &= \frac{1}{2}\sum_{j=0}^{n}\frac{(j-np)^2}{(j+k)^2}\binom{n}{j}p^j(1-p)^{n-j} \\ &\leq \frac{1}{2}\sum_{j=0}^{n}\frac{(j-np)^2}{(j+2)(j+1)}\binom{n}{j}p^j(1-p)^{n-j} \\ &= \frac{1}{2}\sum_{j=0}^{n}\frac{(j-np)^2}{(n+2)(n+1)p^2}\binom{n+2}{j+2}p^{j+2}(1-p)^{n-j} \\ &\leq \frac{\mathbb{E}\left[(V-np)^2\right]}{2(n+2)(n+1)p^2} \\ &= \frac{(n+2)p(1-p)+4p^2}{2(n+2)(n+1)p^2} \\ &\leq \frac{(n+2)p}{2(n+2)(n+1)p^2} \\ &\leq \frac{1}{2np} \leq \frac{1}{np+k}\end{aligned}$$

for $n \geq 4, k \geq 2$ using $k \leq np$ in the last step.

If $c = W + k \geq k$ and $np \leq k$, so

$$\begin{aligned}\mathbb{E}\left[\frac{(W-np)^2}{2c^2}\right] &\leq \mathbb{E}\left[\frac{(W-np)^2}{2k^2}\right] \\ &= \frac{np(1-p)}{2k^2} \\ &\leq \frac{1}{2k} \leq \frac{1}{np+k}\end{aligned}$$

Putting this together,

$$\begin{aligned}&\left|\mathbb{E}\left[\log\left(\frac{W+k}{np+k}\right)\right]\right| \\ &\leq \max\left\{\left|\mathbb{E}\left[\frac{(W-np)^2}{2c^2}\right]\right|, \left|\mathbb{E}\left[\frac{(W-np)^2}{2(W+k)^2}\right]\right|\right\} \\ &= \max\left\{\frac{1}{2(np+k)}, \frac{1}{np+k}\right\} = \frac{1}{np+k}\end{aligned} \tag{4.57}$$

□

Lemma 4.6.4. *Assume $W_n - k \sim$ Binomial$(n-k-1, p)$ and $k \geq \frac{p}{1-p}$. Then*

$$|\mathbb{E}[\log(W_n)] - \log(np)| \leq \frac{1}{k} + \frac{k}{np} \tag{4.58}$$

and

$$|\mathbb{E}[\psi(W_n)] - \log(np)| \leq \frac{2}{k} + \frac{k}{np} \tag{4.59}$$

Proof. Using the triangle inequality,

$$\begin{aligned}&|\mathbb{E}[\psi(W_n)] - \log(np)| \\ &\leq \mathbb{E}[|\psi(W_n) - \log W_n|] + |\mathbb{E}[\log(W_n)] - \log(np)|.\end{aligned}$$

Using lemma 4.6.3, and the fact that $|\log(w)| \leq w - 1$ for $w > 1$, we have that

$$\begin{aligned}
&|\mathbb{E}[\log(W_n)] - \log(np)| \\
&\leq |\mathbb{E}[\log(W_n)] - \log((n-k-1)p+k)| \\
&\quad + |\log((n-k-1)p+k) - \log(np)| \\
&= \left|\mathbb{E}\left[\log\left(\frac{W_n}{(n-k-1)p+k}\right)\right]\right| \\
&\quad + \log\left(\frac{(n-k-1)p+k}{np}\right) \\
&\leq \frac{1}{(n-k-1)p+k} + \frac{(n-k-1)p+k}{np} - 1 \\
&\leq \frac{1}{k} + \frac{k}{np}.
\end{aligned}$$

Because $k \geq \frac{p}{1-p}$, $(n-k-1)p+k \geq np$.

Because $|\psi(w) - \log(w)| < \frac{1}{w}$ for $w > 0$ and $W_n \geq k$, $\mathbb{E}[|\psi(W_n) - \log W_n|] < \mathbb{E}\left[\frac{1}{W_n}\right] \leq \frac{1}{k}$. So,

$$|\mathbb{E}[\psi(W_n)] - \log(np)| \leq \frac{2}{k} + \frac{k}{np}.$$

□

Lemma 4.6.5. *Let V be a d-dimensional random variable on the probability space $(\mathcal{V}, \mathcal{B}_\mathcal{V}, P)$ with $\mathcal{V} = \prod_{i\in I} \mathcal{V} \subseteq \mathbb{R}^d$ where $I = \{1, \ldots, d\}$ and for nonempty $J \subseteq I$, let $P_J = P_{V_i : i\in J}$. Assume that the support of P is $\mathcal{V}$ and that for any nonempty $J \subseteq I$, the set*

$$D_J = \left\{ w \in \prod_{i\in J} \mathcal{V}_i : P_J(\{w\}) > 0 \right\}$$

is countable and nowhere dense in $\prod_{i\in J} \mathcal{V}_i$. Let $v_1, \ldots, v_n \sim P$ be an independent sample in $\mathcal{V}$, and for a point $v \in \mathcal{V}$, define $\tilde{k}(v) = |\{v_i : \|v - v_i\|_\infty \leq \rho_v\}|$ where ρ_v is the distance to the kth nearest neighbor to v in the sample. If $\frac{k}{n} \to 0$, and $n \to \infty$ then

$$\tilde{k}(V) \to k \text{ almost surely}$$

given that $V \in C \equiv \{v \in \mathcal{V} : P(\{v\}) = 0\}$

Proof. For each $J \subseteq I$, D_J is countable, so we can index it with the positive integers. Using contradiction, assume that some ordering is a Cauchy sequence; that is, for every $\epsilon > 0$, there is a positive integer N such that for all integers $l, m \geq N$, $\|a_l - a_m\|_\infty < \epsilon$. But, all Cauchy sequences converge in the complete metric space ([76, Theorem 3.11]), $(\mathbb{R}^d, \ell_\infty)$, so for some a, $a_i \to a$ as $i \to \infty$, a contradiction since D_j is nowhere dense in W. Thus, for each J, there is a $\zeta_J > 0$ such that for any two points, $a_l, a_m \in D_J$, $\|a_l - a_m\|_\infty \geq \zeta_J$. I is finite, so $\zeta \equiv \frac{\min_{J \subseteq I}\{\zeta_J\}}{3}$ exists.

In $(\mathbb{R}^d, \ell_\infty)$, if $\|a - b\|_\infty = \|a - c\|_\infty$, then there must be at least one coordinate, i, such that $a(i) - b(i) = a(i) - c(i) \equiv r$ (where the vectors are function mapping the coordinate(s) to its coordinate value(s)) and for all other coordinates, $j \neq i$, $a(j) - b(j), a(j) - c(j) \leq r$. This can only happen when $a(i), b(i), c(i) \in D_i$; they have a positive point mass so ties are possible. Consider a case where there are discrete points, $P_i(\{w(i)\}) > 0$ for some coordinates, $i \in J$, but will not have any ties in distance. Suppose $A = \{(a(i) : i \in I)\}$ is a subset in the support of P with positive measure such that the marginal distribution on A is discrete for the coordinates in J and continuous for coordinate in $I \backslash J$; that is, $P_i(\{a(i)\}) > 0$ when $i \in J$ and $P_i(\{a(i)\}) = 0$ when $i \in I \backslash J$. Assume that for some point $(b(i) : i \in J)$ in a subspace of A, $P_J((b_i : i \in J)) > 0$, then the subset of A restricted to equal $(b(i) : i \in J)$ on J,

$$B \equiv \{(a(i) : i \in I) \in A : a(i) = b(j), j \in J\},$$

also has a positive probability. If the random sample has values $v_m, v_l \in B$ and another arbitrary point $b \in B$, then

$$\mathbb{P}\left(\|v_m - b\|_\infty = \|v_l - b\|_\infty\right) = 0.$$

This is because the scalar values of $v_m(i), v_l(i)$ and $b(i)$ are equal for $i \in J$ while

for $i \in I \backslash J$, $v_m(i), v_l(i)$ and $b(i)$ are from a continuous distribution, so equal with probability zero with each positive scalar distances. Further, if there are at least k sample points in B, each will have a distinct ℓ_∞-distance to b for the same reason. Thus, $\tilde{k}(b) = k$ with probability one.

Generalizing on this point, let $v \in C$ and let $J = \{i \in I : P_i(\{v(i)\}) = 0\}$. Define $B_J(v, \delta) \subseteq \mathcal{V}$ to be the Cartesian product of $[v(i) - \delta, v(i) + \delta]$ for $i \in J$ and $\{v(i)\}$ for $i \in I \backslash J$,

$$B_J(v, \delta) = \prod_{i \in J} [v(i) - \delta, v(i) + \delta] \times \prod_{i \in I \backslash J} \{v(i)\}.$$

$\mathcal{V}_J$ may have positive point masses among continuous points. Because D_J is nowhere dense in $\mathcal{V}_J$, there is $\delta_v > 0$ such that $D_J \cap \pi_J(B_J(v, \delta_v)) = \varnothing$ (where π_J is the projection onto J) and $P_J(B_J(v, \delta_v)) > 0$. Notice that if there are more than k sample points in $B(v, \delta_v)$ then $\tilde{k}(v) = k$.

Let $W = \{v \in C : \delta_v \geq \zeta\}$ and fix $w \in W$. Let $\epsilon \in [\zeta, 0)$ and ρ_v be the ℓ_∞-distance from v to its kNN in the sample $v_1, \ldots, v_n$. Choose N large enough so that for all $n \geq N$, $\frac{k}{n} \leq P(B_J(v, \epsilon))$. Then using Chernoff's bound,

$$\begin{aligned} \mathbb{P}(\rho_v > \epsilon) &= \mathbb{P}(\text{Binomial}(n, P(v, \epsilon)) \leq k) \\ &\leq \exp\left\{ -\left(\frac{1}{2} n P(v, \epsilon) - k \right) \right\}. \end{aligned}$$

So, $\sum_{n=1}^{\infty} \mathbb{P}(\rho_v > \epsilon) < \infty$. Using the Borel-Cantelli lemma, [46, Lemma 2.2.4], $\rho_v \to 0$ almost surely as $n \to \infty$.

Notice that

$$\mathbb{P}(\rho_V > \epsilon | V \in W) = \int_W \mathbb{P}(\rho_v > \epsilon) dP(v).$$

Using the Lebesgue dominated convergence theorem, with the fact that for each $n, \mathbb{P}(\rho_v > \epsilon) \leq 1$ for all $v \in W$ and $\mathbb{P}(\rho_v > \epsilon) \to 0$ almost surely, we have $\mathbb{P}(\rho_V > \epsilon | V \in W) \to 0$ almost surely as $n \to \infty$. Then $\tilde{k}(V) \to k$ given that $V \in W$ almost surely as $n \to \infty$.

Consider

$$C\backslash W = \{v \in C : \delta_v < \zeta\}.$$

For each $v \in C\backslash W$, there must be $J \subseteq I$ such that

$$D \equiv (D_J \times \{v(I\backslash J)\}) \cap B_J(v, \zeta) \neq \varnothing.$$

There may be points $x \in D$ such that $P(\{x\}) = 0$. Notice that

$$C\backslash W = \bigcup_{x \in D} B(x, \zeta).$$

Similarly, for each $x \in D$, there is $J \subseteq I$ such that $x(J) \in D_J$. Because $D \subseteq \bigcup_{J \subseteq I}(D_J \times D_{I\backslash J})$, D is countable. By choice of ζ, for every two points $a, b \in D$, $\|a - b\|_\infty > \zeta$, so $i \neq j, B(x_i, \zeta) \cap B(x_j, \zeta) = \varnothing$.. With both of these,

$$P\left(\bigcup_{x \in D} B(x, \zeta)\right) = \sum_{i=1}^{\infty} P\left(B(x_i, \zeta)\right).$$

For $x_i \in D$, for all $v \in B(x_i, \zeta)$, there is no J, such that $v(J) \in D_J$ because of choice of ζ. Stated differently, for each $J \subseteq I$, $P(\{v(J)\}) = 0$. Consequently, there can be no ties in distance to points other than x_i. Let $K_{x_i} = \{v \in B(x_i, \zeta)\backslash\{x_i\} : x_i \in B(v, \rho_v)\}$ Using [75, Corollary 6.1], $|K_{x_i}| \leq k\gamma_d$ where γ_d is a function of only the dimension d. Let $p_i = P(B(x_i, \zeta)\backslash\{x_i\})$, then

$$\begin{aligned}
&\mathbb{P}\left(\tilde{k}(v) > k : v \in B(x_i, \zeta)\backslash\{x_i\}\right) \\
&\leq \mathbb{P}(x_i \in B(v, \rho_v)) \\
&= \mathbb{P}(v \in K_{x_i}).
\end{aligned}$$

This probability depends on the number of sample points that fall into $B(x_i, \zeta)$.

Looking at the random variable and using Chernoff,

$$\begin{aligned}
&\mathbb{P}(\tilde{k}(V) > k | V \in B(x_i, \zeta) \backslash \{x_i\}) \\
&\leq \mathbb{P}(V \in K_{x_i} | V \in B(x_i, \zeta) \backslash \{x_i\}) \\
&= \mathbb{P}\left(\text{Binomial}(n, p_i) \leq k\gamma_d\right) \\
&\leq \exp\left\{-\left(\frac{1}{2}np_i - k\gamma_d\right)\right\}.
\end{aligned}$$

So, $\sum_{n=1}^{\infty} \mathbb{P}(\tilde{k}(V) > k | V \in B(x_i, \zeta) \backslash \{x_i\}) < \infty$. Using the Borel-Cantelli lemma, [46, Lemma 2.2.4], $\tilde{k}(V) \to k$ given that $V \in C \backslash W$ almost surely as $n \to \infty$. □

Lemma 4.6.6. *Let $F_\rho(r)$ be the probability that the distance to a point's kNN in a sample of n points is $\rho \leq r$ and let $P_W(r)$ be the probability mass of the ball of radius r centered at the same point. Then*

$$\begin{aligned}
\frac{dF_\rho}{dP_W}(r) = \frac{(n-1)!}{(k-1)!(n-k-1)!} \times \\
[P_W(r)]^{k-1} [1 - P_W(r)]^{n-k-1}.
\end{aligned} \tag{4.60}$$

Proof. Let $\rho_1, \ldots, \rho_{n-1}$ be the ordered distances from the point of interest. The probability that kth largest distance is at least r is

$$\begin{aligned}
&\mathbb{P}(\rho_k \leq r) \\
&= \mathbb{P}(I(\rho_i \leq r) \geq k) \\
&= \sum_{j=k}^{n-1} \mathbb{P}(I(\rho_i \leq r) = j) \\
&= \sum_{j=k}^{n-1} \binom{n-1}{j} [P_W(r)]^j [1 - P_W(r)]^{n-j-1}.
\end{aligned}$$

Taking the derivative with respect to $P_w(r) \equiv p$,

$$\begin{aligned}
&\frac{dF_\rho}{dp} \\
&= \sum_{j=k}^{n-1} \binom{n-1}{j} \frac{d}{dp} \left[p^j (1-p)^{n-j-1}\right] \\
&= \sum_{j=k}^{n-1} \binom{n-1}{j} [j p^{j-1} (1-p)^{n-j-1} \\
&\quad - p^j (n-j-1)(1-p)^{n-j-2}] \\
&= \sum_{j=k}^{n-1} \frac{(n-1)!}{(j-1)!(n-j-1)} p^{j-1} (1-p)^{n-j-1} \\
&\quad - \sum_{j=k}^{n-1} \frac{(n-1)!}{j!(n-j-2)} p^j (1-p)^{n-j-2} \\
&= \frac{(n-1)!}{(k-1)!(n-k-1)!} p^{k-1} (1-p)^{n-k-1}.
\end{aligned}$$

The last equality follows from realizing that all terms cancel except for $j = k$ in the first term. □

Proof of Theorem 4.2.1. We construct the product measure, μ by looking at the scalar coordinates of $V \equiv (V_1, \ldots, V_d)$ over its product space, $\mathcal{V} \equiv \mathcal{V}_1 \times \mathcal{V}_2 \times \cdots \times \mathcal{V}_d$. If $\mathcal{V}_i$ is not a subset of $\mathbb{R}$, V_i is categorical and we use a zero-one distance metric. So that we can work exclusively in $\mathbb{R}^c$ for some positive integer c, we create dummy indicators for all categories except one; this preserves the ℓ_∞ metric for categorical variables. Recall that the marginal measure for any scalar coordinate is $P_{V_i}(A) = P_V(\mathcal{V}_1 \cdots \times \mathcal{V}_{i-1} \times A \times \mathcal{V}_{i+1} \times \cdots \times \mathcal{V}_d)$ where $A \subseteq \mathcal{V}_i$. For each $i = 1, \ldots, d$, redefine $\mathcal{V}_i$ by restricting it to the support of P_{V_i} and $\mathcal{B}_{\mathcal{V}_i}$ the corresponding σ-algebra. Partition $\mathcal{V}_i$ into its discrete and continuous parts. For a set A contained within the support of a random variable, U, let $C_U(A) = \{x \in A : P_U(x) = 0\}$ be the continuous partition and $D_U(A) = \{x \in A : P_U(x) > 0\}$, which is countable by assumption. Clearly $C_U(A) \cup D_U(A) = A$ and $C_U(A) \cap D_U(A) = \varnothing$ for all random variables U. Let

λ be the Lebesgue measure and ν be the counting measure. Define the measure $\mu_i : \mathcal{B}_{\mathcal{V}_i} \to [0, \infty)$ to be $\lambda + \nu_i$, where $\nu_i(C_{V_i}(\mathcal{V}_i)) = 0$ and the counting measure on $D_{V_i}(\mathcal{V}_i)$, $\nu_i(D_{V_i}(\mathcal{V}_i)) = \nu(D_{V_i}(\mathcal{V}_i))$. It is easy to see that μ_i is a well-defined measure on the measurable space, $(\mathcal{V}_i, \mathcal{B}_{\mathcal{V}_i})$ because both the counting measure and Lebesgue measures are well-defined, as is their sum. Define a measure $\mu : \mathcal{B}_{\mathcal{V}} \to \mathbb{R}$ as the product measure, $\mu = \mu_1 \times \mu_2 \times \cdots \times \mu_d$.

With the construction complete, we now show that $P_V \ll \mu$. We begin by showing that for each coordinate, $j = 1, \ldots, d$, $P_{\mathcal{V}_j} \ll \mu_j$. Let $j = 1, \ldots, d$ and $A \in \mathcal{B}_{\mathcal{V}_i}$ with $\mu_j(A) = 0$. Consider the continuous and discrete partitions, $C_{V_j}(A)$ and $D_{V_j}(A)$, respectively. By definition, $\lambda(C_{V_j}(A)) + \nu_j(D_{V_j}(A)) = 0$ so $\lambda(C_{V_j}(A)) = 0$ and $\nu_j(D_{V_j}(A)) = 0$. If the coordinate project for j has a nonempty continuous partition, then $P_{V_j} \ll \lambda$ on $C_{V_j}(\mathcal{V}_j)$, so $P_{V_j}(C_{V_j}(A)) = 0$. Also, $0 = \nu_j(D_{V_j}(A)) = \nu(D_{V_j}(A))$, so $D_{V_j}(A) = \varnothing$, so $P_{V_j}(D_{V_j}(A)) = 0$. Then $P_{V_j}(A) = P_{V_j}(C_{V_j}(A)) + P_{V_j}(D_{V_j}(A)) = 0$

Proceeding by mathematical induction, we already have $P_{V_1} \ll \mu_1$. Assume that $P_{V_1 \ldots V_j} \ll \mu_1 \times \cdots \times \mu_j \equiv \prod_{i=1}^{j} \mu_i$ and that for some $A \in \mathcal{B}_{\mathcal{V}_1 \ldots \mathcal{V}_j \mathcal{V}_{j+1}}$ (the product σ-algebra) $(\prod_{i=1}^{j+1} \mu_i)(A) = 0$. Let $A_{v_1,\ldots,v_j} = \{v_{j+1} : (v_1, \ldots, v_j, v_{j+1}) \in A\}$ and $A_{v_{j+1}} = \{(v_1, \ldots, v_j) : (v_1, \ldots, v_j, v_{j+1}) \in A\}$. Let $A_1 = \mathcal{V}_1 \times \cdots \times \mathcal{V}_j \left\{v_{j+1} : P_{V_1 \ldots V_j}(A_{v_{j+1}}) > 0\right\}$ and $A_2 = \left\{(v_1, \ldots, v_j) : P_{V_{j+1}}(A_{v_1,\ldots,v_j}) > 0\right\} \times \mathcal{V}_{j+1}$.

Using Fubini's theorem,

$$\begin{aligned} 0 &= \left(\prod_{i=1}^{j+1} \mu_i\right)(A) \\ &= \left(\prod_{i=1}^{j} \mu_i \times \mu_{j+1}\right)(A) \\ &= \int_{\mathcal{V}_{j+1}} \left(\prod_{i=1}^{j} \mu_i\right)(A_{v_{j+1}}) d\mu_{j+1}(v_{j+1}). \end{aligned}$$

Using [46, Lemma 1.3.8], $f \geq 0, \int f d\mu = 0 \Rightarrow \mu\{x : f(x) > 0\} = 0$, we must

have

$$\begin{aligned}
0 &= \mu_{j+1}\left(\left\{v_{j+1} : \left(\prod_{i=1}^{j}\mu_i\right) A_{v_{j+1}}) > 0\right\}\right) \\
&= \mu_{j+1}\left(\mathcal{V}_{j+1}\backslash\left\{v_{j+1} : \left(\prod_{i=1}^{j}\mu_i\right)(A_{v_{j+1}}) = 0\right\}\right) \\
&\geq \mu_{j+1}\left(\mathcal{V}_{j+1}\backslash\left\{v_{j+1} : P_{V_1\ldots V_j}(A_{v_{j+1}}) = 0\right\}\right).
\end{aligned}$$

The last inequality follows because $P_{V_1\ldots V_j} \ll \prod_{i=1}^{j}\mu_i$ implies that

$$\left\{v_{j+1} : \left(\prod_{i=1}^{j}\mu_i\right)(A_{v_{j+1}}) = 0\right\} \subseteq \left\{v_{j+1} : P_{V_1\ldots V_j}(A_{v_{j+1}}) = 0\right\}. \tag{4.61}$$

Then $\mu_{j+1}\left(\mathcal{V}_{j+1}\backslash\left\{v_{j+1} : P_{V_1\ldots V_j}(A_{v_{j+1}}) = 0\right\}\right) = 0$. But, $P_{V_{j+1}} \ll \mu_{j+1}$ implies that

$$\begin{aligned}
0 &= P_{V_{j+1}}\left(\mathcal{V}_{j+1}\backslash\left\{v_{j+1} : P_{V_1\ldots V_j}(A_{v_{j+1}}) = 0\right\}\right) \\
&= P_{V_{j+1}}\left(\left\{v_{j+1} : P_{V_1\ldots V_j}(A_{v_{j+1}}) > 0\right\}\right) \\
&= P_{V_1\ldots V_j V_{j+1}}(A_1).
\end{aligned}$$

Using the same procedure but switching $\prod_{i=1}^{j}\mu_i$ and μ_{j+1} and correspondingly, switching $P_{V_1\ldots V_j}$ and $P_{V_{j+1}}$, it is easy to show that

$$\begin{aligned}
0 &= P_{V_1\ldots V_j}\left(\left\{(v_1,\ldots,v_j) : P_{V_{j+1}}(A_{v_1,\ldots,v_j}) > 0\right\}\right) \\
&= P_{V_1\ldots V_j V_{j+1}}(A_2).
\end{aligned}$$

Now, consider the set of points $(v_1,\ldots,v_j,v_{j+1})$ such that each coordinate satisfies $P_{V_{j+1}}(A_{v_1,\ldots,v_j}) = 0$ and $P_{V_1\ldots V_j}(A_{v_{j+1}}) = 0$; call this set, A_3. Showing that

$P(A_3) = 0$, consider the set of points, $(a_1, \ldots a_d) \in B \subseteq \mathcal{V}$ such that

$$P_{V_{j+1}\ldots V_d}\left(\left[A \times \prod_{i=j+2}^{d} \mathcal{V}_i\right]_{(a_1,\ldots,a_j)}\right) = 0$$

and

$$P_{V_1\ldots V_j}\left(\left[A \times \prod_{i=j+2}^{d} \mathcal{V}_i\right]_{(a_{j+1},\ldots,a_d)}\right) = 0.$$

Let $(b_1, \ldots, b_d) \in A_3 \times \prod_{i=j+2}^{d} \mathcal{V}_i$. Then

$$\begin{aligned}
&P_{V_{j+1}\ldots V_d}\left(\left[A \times \prod_{i=j+2}^{d} \mathcal{V}_i\right]_{(b_1,\ldots,b_j)}\right) \\
&= P_{V_{j+1}\ldots V_d}\left(A_{(b_1,\ldots,b_j)} \times \prod_{i=j+2}^{d} \mathcal{V}_i\right) \\
&= P_{V_{j+1}}\left(A_{(b_1,\ldots,b_j)}\right) = 0
\end{aligned}$$

and

$$\begin{aligned}
&P_{V_1\ldots V_j}\left(\left[A \times \prod_{i=j+2}^{d} \mathcal{V}_i\right]_{(b_{j+1},\ldots,b_d)}\right) \\
&= P_{V_1\ldots V_j}\left(A_{b_{j+1}}\right) = 0.
\end{aligned}$$

Then

$$A_3 \times \prod_{i=j+2}^{d} \mathcal{V}_i \subseteq B.$$

Because P_V is non-singular, $P_V(B) = 0$, so

$$P_V\left(A_3 \times \prod_{i=j+2}^{d} \mathcal{V}_i\right) = P_{V_1\ldots V_j V_{j+1}}(A_3) = 0.$$

Now, $A \subseteq A_1 \cup A_2 \cup A_3$ implies that

$$\begin{aligned} & P_{V_1 \ldots V_j V_{j+1}}(A) \\ & \leq P_{V_1 \ldots V_j V_{j+1}}(A_1 \cup A_2 \cup A_3) \\ & \leq P_{V_1 \ldots V_j V_{j+1}}(A_1) + P_{V_1 \ldots V_j V_{j+1}}(A_2) \\ & \quad + P_{V_1 \ldots V_j V_{j+1}}(A_3) \\ & = 0, \end{aligned}$$

so $P_{V_1 \ldots V_j V_{j+1}}(A) = 0$. Thus, by mathematical induction, for any positive integer, d, we have that $P_V \ll \mu$. □

Lemma 4.6.7. *Let μ and ν be nonsingular probability measures on $(\mathbb{R}^d, \mathcal{B})$ such that $\nu \ll \mu$ and assume $\{x : \mu(\{x\}) > 0\}$ is nowhere dense in $\mathbb{R}^d$. Let $B(x, r)$ be a ball of radius r centered at x. If $\mu(\{x\}) > 0$ then*

$$\frac{d\nu}{d\mu}(x) = \frac{\nu(\{x\})}{\mu(\{x\})}$$

otherwise

$$\frac{d\nu}{d\mu}(x) = \lim_{r \to 0} \frac{\nu(B(x, r))}{\mu(B(x, r))}. \tag{4.62}$$

Proof. If $\mu(\{x\}) > 0$, then $\frac{d\nu}{d\mu}(x) = \frac{\nu(\{x\})}{\mu(\{x\})}$:

$$\int_{\{x\}} \frac{\nu(\{x\})}{\mu(\{x\})} d\mu = \frac{\nu(\{x\})}{\mu(\{x\})} \mu(\{x\}) = \nu(\{x\}).$$

If $\mu(\{x\}) = 0$ and in the support of μ, there must be some $\delta > 0$ such that for every $y \in B(x, \delta), \mu(\{y\}) = 0$ because $\{x : \mu(\{x\}) > 0\}$ is nowhere dense in $\mathbb{R}^d$ and $\mu(B(x, \delta)) > 0$. Notice that some coordinates of $x = (x_1, \ldots, x_d)$ may be discrete but there must be at least one continuous coordinate in order for $\mu(\{x\}) = 0$. Let I_{cont} be the index of continuous coordinates of x and I_{disc} be the index of discrete coordinates of x. Using the proof of lemma 4.2.1, each coordinate of I_{cont} will be dominated by

the Lebesgue measure within $B(x,\delta)$. Again, because $\{x : \mu(\{x\}) > 0\}$ is nowhere dense in $\mathbb{R}^d$,

$$\delta_{\text{disc}} \equiv \min_{y \in \text{Supp}(\mu)} \left\{ \|x_{\text{disc}} - y_{\text{disc}}\|_{\ell_\infty} : x_{\text{disc}} \neq y_{\text{disc}} \right\} > 0$$

where $z_{\text{disc}} \equiv (z_i : I_{\text{disc}})$ for $z = x, y$. If $\delta > \delta_{\text{disc}}$, then redefine $\delta = \delta_{\text{disc}}$. Now x_{disc} is constant within $B(x,\delta)$ and homeomorphic to a subset of $\mathbb{R}^a$ for some integer $a \leq d$ with the corresponding Lebesgue measure. Then $\mu \ll \lambda$ on $B(x,\delta)$ where λ is Lebesgue on the support of μ and zero otherwise so that $\lambda \ll \mu$ as well. Using [76, Theorem 7.8],

$$\frac{d\nu}{d\lambda}(x) = \lim_{r \to 0} \frac{\nu(B(x,r))}{\lambda(B(x,r))}$$

and

$$\frac{d\mu}{d\lambda}(x) = \lim_{r \to 0} \frac{\mu(B(x,r))}{\lambda(B(x,r))}.$$

Notice that $\mu \ll \lambda$ and $\lambda \ll \mu \Rightarrow \frac{d\mu}{d\lambda}(x) > 0$. Then

$$\begin{aligned}
\frac{d\nu}{d\mu}(x) &= \left(\frac{d\nu}{d\lambda} \frac{d\lambda}{d\mu} \right)(x) \\
&= \left[\frac{d\nu}{d\lambda} \left(\frac{d\lambda}{d\mu} \right)^{-1} \right](x) \\
&= \left(\lim_{r \to 0} \frac{\nu(B(x,r))}{\lambda(B(x,r))} \right) \left(\lim_{r \to 0} \frac{\mu(B(x,r))}{\lambda(B(x,r))} \right)^{-1} \\
&= \lim_{r \to 0} \frac{\nu(B(x,r))}{\mu(B(x,r))}.
\end{aligned}$$

□

Lemma 4.6.8. *Assume $0 < f \leq C$, for some $C > 0$ if $P_{XYZ}(\{(x,y,z)\}) > 0$ then*

$$f(x,y,z) = \frac{P_{XYZ}(\{(x,y,z)\}) P_Z(\{z\})}{P_{XZ}(\{(x,z)\}) P_{YZ}(\{(y,z)\})}$$

otherwise

$$\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} \to \frac{dP_{XY|Z}}{d(P_{X|Z} \times P_{Y|Z})} \tag{4.63}$$

(converges pointwise) as $r \to 0$ *and*

$$\frac{P_{XYZ}(r)P_Z(r)}{P_{XZ}(r)P_{YZ}(r)} \leq C \tag{4.64}$$

almost everywhere $[P_{X|Z} \times P_{Y|Z}]$.

Proof. From lemma 4.2.1, for a random variables, U and V define $\mu_U \times \mu_V = \mu_{UV}$. Based on definitions from [45, §7.1 and §7.2], define $p_{UV} = \frac{dP_{UV}}{d\mu_{UV}}$, $p_V = \frac{dP_V}{d\mu_V}$, and $p_{U|V} = \frac{p_{UV}}{p_V}$. Note that μ_{UV} is not a probability measure, but for brevity, we define $\mu_{U|V}(A|v) = \mu_U(A)$ for A in the support of U and v in the support of V so that $P_{U|V}$ and $\mu_{U|V}$ have the same support.

From lemma 4.2.1, $P_{XY|Z} \ll \mu_{XY|Z}$. Because $P_{X|Z}$ has the same support as $\mu_{X|Z}$, $\mu_{X|Z} \ll P_{X|Z}$; similarly, $P_{Y|Z}$ has the same support as $\mu_{Y|Z}$, so $\mu_{Y|Z} \ll \mathbb{P}_{Y|Z}$. Using a proof similar to that of lemma 4.2.1, $\mu_{X|Z} \times \mu_{Y|Z} \ll P_{X|Z} \times P_{Y|Z}$. But, $\mu_{XY|Z} = \mu_{X|Z} \times \mu_{Y|Z}$ because it is a product measure.

Using properties of RN derivatives so that

$$\begin{aligned}
&\frac{dP_{XY|Z}}{d(P_{X|Z}\times P_{Y|Z})}\\
&=\frac{dP_{XY|Z}}{d(\mu_{X|Z}\times\mu_{Y|Z})}\frac{d(\mu_{X|Z}\times\mu_{Y|Z})}{d(P_{X|Z}\times P_{Y|Z})}\\
&=\frac{dP_{XY|Z}}{d\mu_{XY|Z}}\left[\frac{d(P_{X|Z}\times P_{Y|Z})}{d(\mu_{X|Z}\times\mu_{Y|Z})}\right]^{-1}\\
&=\frac{dP_{XY|Z}/d\mu_{XY|Z}}{d(P_{X|Z}\times P_{Y|Z})/d(\mu_{X|Z}\times\mu_{Y|Z})}\\
&=\frac{dP_{XY|Z}/d\mu_{XY|Z}}{(dP_{X|Z}/d\mu_{X|Z})(dP_{Y|Z}/d\mu_{Y|Z})}\\
&=\frac{\frac{dP_{XYZ}}{d\mu_{XYZ}}/\frac{dP_Z}{d\mu_Z}}{\left(\frac{dP_{XZ}}{d\mu_{XZ}}/\frac{dP_Z}{d\mu_Z}\right)\left(\frac{dP_{YZ}}{d\mu_{YZ}}/\frac{dP_Z}{d\mu_Z}\right)}\\
&=\frac{\frac{dP_{XYZ}}{d\mu_{XYZ}}\frac{dP_Z}{d\mu_Z}}{\frac{dP_{XZ}}{d\mu_{XZ}}\frac{dP_{YZ}}{d\mu_{YZ}}}\\
&=\frac{d(P_{XYZ}\times P_Z)/d(\mu_{XYZ}\times\mu_Z)}{d(P_{XZ}\times P_{YZ})/d(\mu_{XZ}\times\mu_{YZ})}\\
&=\frac{d(P_{XYZ}\times P_Z)}{d(P_{XZ}\times P_{YZ})}.
\end{aligned}$$

Applying lemma 4.6.7 completes the first claim.

Second, note that $g\equiv\frac{d\nu}{d\mu}\leq C$ implies that for any set A such that $\mu(A)>0$, $\frac{\nu(A)}{\mu(A)}\leq C$. To see this, $\nu(A)=\int_A g d\mu\leq\int_A C d\mu=C\mu(A)$. So, the second claim holds as well. □

Lemma 4.6.9. *Assume* $W_{n,r}-k\sim$ *Binomial* $\left(n-k-1,\frac{q(r)-p(r)}{1-p(r)}\right)$ *where* $p(r),q(r)$ *are probabilities, and for all* r, $p(r)\leq q(r)$. *Then*

$$\left|\int_0^\infty \mathbb{E}\left[\psi(W_{n,r})\right]-\log(nq(r))dF_\rho(r)\right|<\frac{3}{k-1} \tag{4.65}$$

and

$$\left| \int_0^\infty \mathbb{E}\left[\log(W_{n,r})\right] - \log(nq(r)) dF_\rho(r) \right| < \frac{2}{k-1}. \tag{4.66}$$

Proof. We suppress the arguments/subscripts, r and n for brevity through out this proof. Using the triangle inequality, the fact that $|\psi(w) - \log(w)| \leq \frac{1}{w}$,

$$\begin{aligned}
&\left|\mathbb{E}\left[\psi(W_{n,r})\right] - \log(nq(r))\right| \\
&\equiv \left|\mathbb{E}\left[\psi(W)\right] - \log(nq)\right| \\
&\leq \left|\mathbb{E}\left[\psi(W)\right] - \mathbb{E}\left[\log(W_n)\right]\right| \\
&\quad + \left|\mathbb{E}\left[\log(W)\right] - \log\left[(n-k-1)\left(\frac{q-p}{1-p}\right) + k\right]\right| \\
&\quad + \left|\log\left[(n-k-1)\left(\frac{q-p}{1-p}\right) + k\right] - \log(nq)\right| \\
&\leq \mathbb{E}\left[\frac{1}{W}\right] + \left|\mathbb{E}\left[\log\left(\frac{W}{(n-k-1)\left(\frac{q-p}{1-p}\right) + k}\right)\right]\right| \\
&\quad + \log\left(\frac{(n-k-1)\left(\frac{q-p}{1-p}\right) + k}{nq}\right) \\
&\leq \frac{1}{k} + \frac{1}{(n-k-1)\left(\frac{q-p}{1-p}\right) + k} + \frac{k}{np} - 1 \\
&\leq \frac{2}{k} + \frac{k}{np} - 1.
\end{aligned}$$

The penultimate step uses $W \geq k$ and proposition 4.6.3 for the first two terms. We show the third term here, again using $\log(w) \leq w - 1$ for $w \geq 0$ and $\left(\frac{p(1-q)}{q(1-p)}\right) \leq 1$:

$$
\begin{aligned}
&\log\left(\frac{(n-k-1)\left(\frac{q-p}{1-p}\right)+k}{nq}\right)\\
&\leq \frac{(n-k-1)\left(\frac{q-p}{1-p}\right)+k}{nq}-1\\
&= \frac{k\left(\frac{1-q}{1-p}\right)+(n-1)\left(\frac{q-p}{1-p}\right)}{nq}-1\\
&= \frac{k}{np}\left(\frac{p(1-q)}{q(1-p)}\right)+\frac{n-1}{n}\left(\frac{q-p}{q(1-p)}\right)-1\\
&= \frac{k}{np}\left(\frac{p(1-q)}{q(1-p)}\right)+\frac{n-1}{n}\left(1-\frac{p(1-q)}{q(1-p)}\right)-1\\
&= \left(\frac{k}{np}-\frac{n-1}{n}\right)\left(\frac{p(1-q)}{q(1-p)}\right)+\frac{1}{n}\\
&\leq \frac{k}{np}-\frac{n-1}{n}+\frac{1}{n}=\frac{k}{np}-1.
\end{aligned}
$$

Putting this all together,

$$
\begin{aligned}
&\left|\int_0^\infty \mathbb{E}\left[\psi(W)\right]-\log(nq)dF_\rho\right|\\
&\leq \int_0^\infty \left|\mathbb{E}\left[\psi(W)\right]-\log(nq)\right|dF_\rho\\
&\leq \int_0^\infty \left(\frac{2}{k}+\frac{k}{np}-1\right)dF_\rho\\
&= \frac{2}{k}+\frac{k}{n}\int_0^\infty \frac{1}{p}dF_\rho-1\\
&= \frac{2}{k}+\frac{k}{n}\left(\frac{n-1}{k-1}\right)-1\\
&\leq \frac{2}{k}+\frac{k}{k-1}-1\leq\frac{3}{k-1}
\end{aligned}
$$

We complete the integration step using lemma 4.6.6 and two beta function iden-

tities,

$$\begin{aligned}
&\int_0^\infty \frac{1}{p} dF_\rho \\
&= \int_0^1 \frac{1}{p} \frac{(n-1)!}{(k-1)!(n-k-1)!} p^{k-1}(1-p)^{n-k-1} dp \\
&= \frac{(n-1)!}{(k-1)!(n-k-1)!} \int_0^1 p^{k-2}(1-p)^{n-k-1} dp \\
&= \frac{(n-1)!}{(k-1)!(n-k-1)!} \frac{(k-2)!(n-k-1)!}{(n-2)!} \\
&= \frac{n-1}{k-1}.
\end{aligned}$$

The second claim follows using a close but simpler argument. □

Chapter 5

Graph Divergence

5.1 Introduction

Researchers frequently use data to glean information on causal pathways. In many settings, collaborations combine regression results with expert understanding to answer a scientific question. This paradigm works well to fine-tune when causal pathways and structure are largely understood; however, this is not always the case. In settings where causal connections between many variables are poorly understood, *causal discovery*, methods for estimating the causal structure underlying a dataset, may be of value for exploratory data analysis and hypothesis generation. For example, with emerging infectious diseases within a populations, it may be helpful to start with causal discovery to attempt to understand transmission pathways.

Interestingly, causal discovery is rarely seen in fields that apply data methods. While the reasons for this are unclear, there may be several contributing factors. Scientific journals tend to prefer articles that point to positive, conclusive results, rather than exploratory outcomes. Consequently, this may also contribute to reduced exposure to and awareness of causal discovery methods with some scientific communities, making it more less likely to be used. Another possible reason could be that most causal discovery methods, other than reproducing kernel Hilbert space estimators [5], are not able to handle mixed data, that is, data including both discrete (or categorical) and continuous variables, or variables that include both discrete and continuous values. Many datasets in applied fields are mixed. Finally, most causal discovery methods offer no way to test the overall accuracy of results or provide confidence sets of graph estimates, analogous to confidence intervals for scalar estimates.

Despite these challenges, much research could stand to benefit from empirically estimating the underlying causal structure of a dataset, or verifying held mental models. Further, causal discovery can aid with hypothesis generation, understanding causal mediation, and policy analysis .

Information theoretic measures such as entropy, mutual information, conditional

mutual information, and Kullback-Leibler (KL) divergence are attractive for this purpose for three primary reasons: First, this class of measures is well defined for general random variables and vectors [45, Lemma 7.3]. Second, other than requiring distributions be absolute continuous with respect to a reference product measure, they place no conditions on distributions to accurately capture either uncertainty, mutual or conditional independence, or similarity. Third, mutual information, conditional mutual information, and KL divergence between a Bayesian factorization and its full, joint distribution can all be decomposed into sums and differences of entropies, providing a way to simplify complex structures.

In this paper, we develop a nonparametric, information-theoretic estimator for quantifying the Kullback-Leibler divergence between a joint distribution and any Bayesian factorization (DAG) of its variables. We show that this estimator is asymptotically Gaussian under some conditions for the ratio of the densities in question. Having an approximate sampling distribution for the estimator allows for tests and confidence sets. Paired with an algorithm to optimize the estimator graph divergence score over the space of DAGs, one could test the accuracy of an optimal DAG (Markov equivalence class) or compute a set of similarly likely Markov-equivalent DAGs. In practice, this method could be helpful for exploratory scientific research.

We organize the paper as follows. Section 5.2 explains the graph divergence metric. Section 5.3 briefly reviews past methods for estimating graph divergence and related metrics. Section 5.4 develops a novel method for estimating graph divergence for continuous random variables which obeys the central limit theorem. Section 5.5 expands on the prior method to work with mixed variables.

5.2 Bayesian Factorization and Divergence

At a high level, information theory provides similarity metrics between probability measures (or distributions). For example, if X and Y are random variables, their mutual information, $I(X;Y)$, quantifies the KL divergence between their joint proba-

bility measure, P_{XY}, and the product of their marginal probability measures, $P_X P_Y$. This case of KL divergence indicates the dependence of X and Y; $I(X;Y) = 0$ if and only if $P_{XY} = P_X P_Y$. Similarly, with a third random variable, Z, the conditional mutual information, $I(X;Y|Z)$, is the KL divergence between $P_{XY|Z}$ and $P_{X|Z}P_{Y|Z}$, and quantifies the conditional mutual information between X and Y given Z. With more variables, there are more possible factorizations indicating the conditional independence relationships between them.

Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional probability space and for $W \subseteq [d] := \{1, \ldots, d\}$ and $V \subseteq W$, let P_W be the marginal probability measure of P with respect W and $P_{W|V}$ be the conditional probability measure of W with respect to V. We assume that all conditional probability measures are regular [27, Section 4.1.3]. Let $\mathcal{G}$ be a Bayesian network [6, Chapter 3], (also called a directed acyclic graph (DAG)) and

$$P_{\mathcal{G}} := \prod_{j=1}^{d} P_{X_j|\mathrm{pa}(j)} \tag{5.1}$$

be a factorization of P according to $\mathcal{G}$ into conditional probability measures where $\mathrm{pa}(j)$ are the parents of X_j in the graph.

Definition 5.2.1. *The graph divergence between P and $P_{\mathcal{G}}$ is*

$$D(\mathcal{G}) := \int_{\mathcal{X}} \log\left(\frac{dP}{dP_{\mathcal{G}}}(x)\right) dP(x) \tag{5.2}$$

where $\frac{dP}{dP_{\mathcal{G}}}$ is the RN derivative of P with respect to $P_{\mathcal{G}}$.

In order for this definition to be well-defined, it is necessary that $\frac{dP}{dP_{\mathcal{G}}}$ exist. A sufficient conditions for this is that the P is non-singular.

Definition 5.2.2. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional probability space with $\mathcal{X} = \prod_{i\in[d]} \mathcal{X}_i$, $\mathcal{X}_J := \prod_{i\in J} \mathcal{X}_i$ and $P_J = P_{\mathcal{X}_J}$ for $J \subseteq [d]$. For $A \subseteq \mathcal{X}$, and $v = (v_i : i \in$*

$J) \in \mathcal{X}_J$, define the section of A with respect to v as

$$A_v := \{(a_i : i \in [d] \backslash J) : (a_i : i \in [d]) \in A, a_i = v_j, i = j \in J\}. \tag{5.3}$$

P is non-singular *if any subset set $A \subseteq \mathcal{X}$ and $J \subseteq [d]$ such that for all $a \in A$, $P_J\left(A_{(a_i:i\in[d]\backslash J)}\right) = 0$ and $P_{[d]\backslash J}\left(A_{(a_i:i\in J)}\right) = 0$ implies $P(A) = 0$.*

This condition ensures that the continuous part of a d-dimensional probability measure does not concentrate on lower dimensional sets. For example, the joint measure of a continuous random variable and a deterministic functional transformation is singular. This is not a problem for the discrete part of the probability measure because its reference measure, the counting measure, is never zero. The following theorem shows this is a sufficient condition for the existence of graph divergence.

Theorem 5.2.1. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional probability space and $P_{\mathcal{G}}$ a Bayesian factorization corresponding to DAG, $\mathcal{G}$. If P is non-singular, then $P \ll P_{\mathcal{G}}$ and the Radon-Nikodym derivative of P with respect to $P_{\mathcal{G}}$, $\frac{dP}{dP_{\mathcal{G}}}$, exists and $D(\mathcal{G}) = \int_{\mathcal{X}} \log \frac{dP}{dP_{\mathcal{G}}} dP$ is well-defined.*

Just as the mutual and conditional mutual information metrics can be written in terms of entropy, so can graph divergence. Let $\text{pa}^*(j) = \text{pa}(j) \cup X_j$ and for $W \in [d]$, let $p_W := \frac{dP_W}{d\mu_W}$ be the density function or RN derivative of P_W with respect to its $|W|$-dimensional reference measure, μ_W [1]. Using properties of conditional probabilities and RN derivatives, the graph divergence can also be expressed in terms of entropy.

Proposition 5.2.2. *If P is a non-singular probability measure and $\mathcal{G}$ is a DAG, then*

$$D(\mathcal{G}) = H(P) - \sum_{j=1}^{d} \left[H(P_{pa^*(j)}) - H(P_{pa(j)})\right] \ . \tag{5.4}$$

[1] see appendix for details on reference measures for mixed probability measures

Proof.

$$D(\mathcal{G}) = \int_{\mathcal{X}} \log \frac{dP}{dP_{\mathcal{G}}} dP \tag{5.5}$$

$$= \int_{\mathcal{X}} \log \left(\frac{dP}{d\left(\prod_{j=1}^{d} P_{X_j|\mathrm{pa}(j)}\right)} \right) dP \tag{5.6}$$

$$= \int_{\mathcal{X}} \log \left(\frac{d\left(P \prod_{j=1}^{d} P_{\mathrm{pa}(j)}\right)}{d\left(\prod_{j=1}^{d} P_{\mathrm{pa}^*(j)}\right)} \right) dP \tag{5.7}$$

$$= \int_{\mathcal{X}} \log \left(\frac{p \prod_{j=1}^{d} p_{\mathrm{pa}(j)}}{\prod_{j=1}^{d} p_{\mathrm{pa}^*(j)}} \right) dP \tag{5.8}$$

$$= \int_{X} \log p - \sum_{j=1}^{d} \left[\log p_{\mathrm{pa}^*(j)} - \log p_{\mathrm{pa}(j)}\right] dP \tag{5.9}$$

$$= H(P) - \sum_{j=1}^{d} \left[H(P_{\mathrm{pa}^*(j)}) - H(P_{\mathrm{pa}(j)})\right] \tag{5.10}$$

where μ is the reference product measure for P. □

Note if a node, say j, has no parents, then $\mathrm{pa}(j) = \varnothing$; in this case $P_{\varnothing} = 1$, vacuously, as expected. Being able to decompose graph divergence into entropy components can provide intuition for the estimator and simplify estimation.

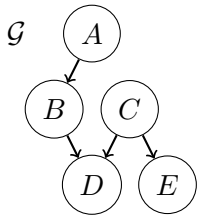
More generally, a DAG provides a succinct way to describe a set of conditional independence relationships for the set of variables that comprise a joint probability measure. For a faithful distribution (see [6, Def. 3.8]), a DAG having zero divergence indicates that its corresponding set of conditional independence relationships are a subset of the set of conditional independence relationships in the probability measure. Because the presence of an edge within a DAG indicates dependence between variables, if one adds an edge to a zero-divergence DAG, the resulting DAG will also have a divergence of zero with respect to the given probability measure. A more in-depth treatment can be found in [6, Chapter 3]. Causal discovery attempts to estimate the minimal (least number of edges) DAG of zero-divergence.

5.2.1 Examples

Let $(\mathcal{X}, \mathcal{B}, P)$ be a 5-dimensional probability space with random variables A, B, C, D and E such that their joint distribution can be factored as

$$\mathbb{P}(A, B, C, D, E) = \mathbb{P}(A)\mathbb{P}(B|A)\mathbb{P}(C)\mathbb{P}(D|BC)\mathbb{P}(E|C). \tag{5.11}$$

This factorization corresponds to the following DAG below:



In this case,

$$\begin{aligned} D(\mathcal{G}) &= \int_{\mathcal{X}} \log\left(\frac{dP_{ABCDE}}{d\left[P_A P_{B|A} P_C P_{D|BC} P_{E|C}\right]}\right) dP_{ABCDE} && (5.12)\\ &= \int_{\mathcal{X}} \log\left(\frac{d\left[P_{ABCDE} P_\varnothing P_A P_\varnothing P_{BC} P_C\right]}{d\left[P_A P_{AB} P_C P_{BCD} P_{CE}\right]}\right) dP_{ABCDE} && (5.13)\\ &= \int_{\mathcal{X}} \log\left(\frac{d\left[P_{ABCDE} P_{BC}\right]}{d\left[P_{AB} P_{BCD} P_{CE}\right]}\right) dP_{ABCDE} && (5.14)\\ &= H(A,B) + H(B,C,D) + H(C,E) - H(A,B,C,D,E) - H(B,C) && (5.15) \end{aligned}$$

The mutual information between two random variables, X and Y, is a special case of KL divergence between the joint distribution and the product of the marginal distributions. It is straightforward to see that

$$I(X;Y) = \int \log\left(\frac{dP_{XY}}{d[P_X P_Y]}\right) dP_{XY} = H(X) + H(Y) - H(X,Y). \tag{5.16}$$

Similarly, conditional mutual information between X and Y conditioning on Z is also a special case of KL divergence between the joint conditional distribution, $P_{XY|Z}$

and the marginal conditional distributions, $P_{X|Z}P_{Y|Z}$,

$$I(X;Y|Z) = \int \log\left(\frac{dP_{XY|Z}}{d[P_{X|Z}P_{Y|Z}]}\right) dP_{XYZ} \tag{5.17}$$

$$= \int \log\left(\frac{d[P_{XYZ}P_Z]}{d[P_{XZ}P_{YZ}]}\right) dP_{XY} \tag{5.18}$$

$$= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)\ . \tag{5.19}$$

5.3 Prior Estimation Methods

The estimation of information theoretic measures for discrete random variables can be based on *plug-in* estimates, substituting the empirical distribution into the defining formulas. [77] showed that the plug-in entropy estimator for discrete random variables with finite range (alphabet) is asymptotically Gaussian. For arbitrary discrete random variables with a countable range, [78] further showed that plug-in estimators for information theoretic measures are universally consistent. Unfortunately, these estimates suffer from finite-sample bias, especially when the range of a random variable is comparatively large [79].

Estimating information theoretic measures for continuous random variables with plug-in estimators can be challenging because it requires estimating the underlying distribution itself. Dmitriev and Tarasenko first proposed such an estimator for functionals [54] for scalar random variables. Darbellay and Vajda [55], in contrast, proposed an estimator mutual information based on frequencies in rectangular partitions. Nearest-neighbor methods of estimating information-theoretic quantities for continuous random variables which evade the step of directly estimating a density go back over thirty years, to [56], which proposed an estimator (KL) of the differential entropy:

$$\widehat{H}_{\mathrm{KL}}(X) = -\psi(k) + \psi(n) + \log c_{d,p} + \frac{d}{n}\sum_{i=1}^{n} \log \rho_{k,i,p} \tag{5.20}$$

where k is the kNN value, n is the sample size, $\rho_{k,i,p}$ is the kNN distance in the ℓ_p

metric, and $c_{c,p} := 2^d \Gamma\left(1+\frac{1}{p}\right)^d / \Gamma\left(1+\frac{d}{p}\right)$, the volume of a d-dimensional, ℓ_p-ball of radius one [57]. [80] shows that, under some conditions, a weighted variation of this estimator is asymptotically Gaussian. [59] builds on the original KL estimator to estimate mutual information based on equation 5.16. With $p = \infty$ in equation 5.20, this method locally estimates $H(X) + H(Y) - H(X,Y)$ for each observation by calculating $\rho_{k,i,\infty}$ for $\widehat{H}(X,Y)$ exactly as $\widehat{H}_{\mathrm{KL}}(X,Y)$. It estimates $\widehat{H}(X)$ and $\widehat{H}(Y)$ differently by reusing the value of $\rho_{k,i,\infty}$ from before and replacing the value of k in these entropy estimates with $n^*_{X,i}$ and $n^*_{Y,i}$, the number of sample points within a radius $\rho_{k,i,\infty}$ ℓ_∞-distance of the ith observation in the lower dimensional X and Y subspaces. These modifications cause $\log c_{d,p} = 0$ and $\log \rho_{k,i,\infty}$ terms to exactly cancel, making the estimator

$$\widehat{I}_{\mathrm{KSG}}(X;Y) = \psi(k) + \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\left[\psi(n^*_{X,i}) + \psi(n^*_{Y,i})\right] \ . \tag{5.21}$$

The original paper did not offer any proofs on consistency or distribution. [58] later shows that both the KL and the KSG estimator are consistent along with their mean-squared error (MSE) rates of convergence. The corresponding assumptions and MSE rates were further developed in reference [81].

This modification is significant for this work because by only requiring counting neighbors within a given radius of each observation, estimation for continuous random variables becomes similar to that of discrete random variable with counting identical values. This insight from [66], on estimating mutual information for mixed data, informed the direction of this work. [67] expanded the previous method to estimate graph divergence. Both papers show show that under some conditions, their estimators are consistent along with convergences. [73] develops an estimator for graph divergence based in kernel density estimation and obeys the central limit theorem.

5.3.1 Discrete Graph Divergence Estimation

Using the standard plug-in estimator for entropy (Antos, Kontoyiannis, 2001) and (Girsanov, 1959), assume $X \sim P$ takes finitely many value and let $X^{(1)}, \ldots, X^{(n)} \sim P$ be an iid sample. Without loss of generality, assume X takes values in $[K]$ and let $p^{(j)} = \mathbb{P}(X = j)$, $k_j := \left|\left\{i \in [n] : X^{(i)} = j\right\}\right|$, and $k^{(i)} := \left|\left\{j \in [n] : X^{(j)} = X^{(i)}\right\}\right|$. Then, defining $0 \log 0 = 0$,

$$\widehat{H}_n(X) = \sum_{j=1}^{K} \widehat{p}_j \log \widehat{p}_j = \sum_{j=1}^{K} \frac{k_j}{n} \log\left(\frac{k_j}{n}\right) \tag{5.22}$$

$$= \frac{1}{n} \sum_{j=1}^{K} k_j \log k_j - k_j \log n = \frac{1}{n} \sum_{i=1}^{n} \log k^{(i)} - \log n \ . \tag{5.23}$$

For $W \in [d]$, let $k_W^{(i)} = \left|\left\{j \in [n] : X_W^{(j)} = X_W^{(i)}\right\}\right|$; that is, the number of sample points equal to $X^{(i)}$ on the coordinates in W. For each $i \in [n]$, letting

$$\xi^{(i)} = \log k^{(i)} - \log n - \sum_{j=1}^{d} \log k^{(i)}_{\mathrm{pa}^*(j)} - k^{(i)}_{\mathrm{pa}(j)}, \tag{5.24}$$

we can estimate graph divergence as

$$\widehat{D}_n(\mathcal{G}) = \widehat{H}_n(P) - \sum_{j=1}^{d} \left[\widehat{H}_n(P_{\mathrm{pa}^*(j)}) - \widehat{H}_n(P_{\mathrm{pa}(j)})\right] \tag{5.25}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \xi^{(i)} \ . \tag{5.26}$$

Theorem 5.3.1. *Let $X \sim P$ be a d-dimensional random variable taking finitely many values, and $\mathcal{G}$ be a DAG on the coordinates of X. If $X(1), \ldots, X^{(n)} \sim P$, then*

$$\sqrt{n}\left(\widehat{D}_n(\mathcal{G}) - D(\mathcal{G})\right) \rightsquigarrow N(0, V(G)) \tag{5.27}$$

where $V(G) = Var(\log f(X))$.

Proof. Because $\widehat{D}(\mathcal{G})$ is a linear combination of entropy estimates, each of which is asymptotically Gaussian, $\widehat{D}(\mathcal{G})$ must also be asymptotically Gaussian as well. □

5.4 Continuous Graph Divergence Estimation

Let $\mathcal{X}, \mathcal{B}, P$ be a d-dimensional probability space such that $\mathcal{X} \subseteq \mathbb{R}$ and $P \ll \lambda$, (P is absolutely continuous with respect to Lebesgue measure). Let $X^n = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\} \sim P$ be an i.i.d. sample from P. For $W \subseteq [d] := \{1, 2, \ldots, d\}$, let $X_W = (X_j : j \in W)$. For $k \in [n-1]$, and $X^{(i)} \in X^n$, let $X^{(i)}_{(k)}$ be the kth nearest neighbor (kNN) to $X^{(i)}$ in $X^n \backslash \left\{X^{(i)}\right\}$ using the ℓ_∞ norm and define $\rho_k^{(i)} := \left\|X^{(i)} - X^{(i)}_{(k)}\right\|_\infty$. When clear, we may drop superscripts. When referring to an observation without an observation number, we use random variable in the superscript.

For $W \subseteq [d]$ and $X, X^{(1)}, \ldots, X^{(n)} \sim P$, define

$$n_W^{(X)} := \left|\left\{X^{(i)} : \left\|X_W - X_W^{(i)}\right\|_\infty < \rho_k^{(X)}\right\}\right|, \tag{5.28}$$

the number of points within an ℓ_∞ radius of $\rho_k^{(X)}$ from X in the dimension of W.

For a Bayesian factorization, $\mathcal{G}$, and each $i \in [n]$, define

$$\xi^{(i)} := \psi(k) - \psi(n+1) - \sum_{j=1}^{d}\left[\psi\left(n^{(i)}_{\mathrm{pa}^*(j)} + 1\right) - \psi\left(n^{(i)}_{\mathrm{pa}(j)} + 1\right)\right] \tag{5.29}$$

where $n^{(i)}_\varnothing = n$. Define the continuous graph divergence estimator as

$$\widehat{D}_n(\mathcal{G}) := \frac{1}{n}\sum_{i=1}^{n}\xi^{(i)}. \tag{5.30}$$

For $W \in [d], r > 0$ and $x \in \mathcal{X}$, define

$$P_W^{(x)}(r) := P_W\left(B_W^{(x)}(r)\right) \tag{5.31}$$

as the W-marginal probability mass within an ℓ_∞ ball (hypercube) of radius r centered at X. No subscript indicates the full, joint distribution of X.

For $r > 0$ and $X \sim P$, consider the ratio

$$f_r(X) = \frac{P^{(X)}(r)}{P_{\mathcal{G}}^{(X)}(r)} = \frac{P^{(X)}(r) \prod_{j=1}^{d} P_{\mathrm{pa}(j)}^{(X)}(r)}{\prod_{j=1}^{d} P_{\mathrm{pa}^*(j)}^{(X)}(r)} \tag{5.32}$$

of probability masses of ℓ_∞ balls, comparing the true density to the Bayesian factorization. Each point estimate approximates this ratio with radius ρ_k:

$$\xi^{(X)} \approx f_{\rho_k}(X) \ . \tag{5.33}$$

Similar to Lebesgue's theorem,

$$f_r(X) \to f(X) \text{ as } r \to 0 \tag{5.34}$$

where $f := \frac{dP}{dP_{\mathcal{G}}}$, the Radon-Nikodym derivative of P with respect to $P_{\mathcal{G}}$.

Lemma 5.4.1. *If $X^{(1)}, \ldots, X^{(n)} \sim P$ is a random sample (superscripts in parenthesis indicate observation number) and*

$$\xi^{(X)} := \psi(k) - \psi(n+1) - \sum_{j=1}^{d} \left[\psi\left(n_{pa^*(j)}^{(X)} + 1 \right) - \psi\left(n_{pa(j)}^{(X)} + 1 \right) \right], \tag{5.35}$$

then

$$\xi^{(X)} = \mathbb{E}\left[\log f_{\rho_k}(X) | X, n_{pa} \right] \tag{5.36}$$

where $n_{pa} := \left\{ n_{pa(j)}, n_{pa^*(j)} : j \in [d] \right\}$.

The proof can be found in appendix 5.8.

In order to control bias, we make two primary assumptions, that the joint distribution and all marginals are sufficiently smooth, and that the it has bounded moments. The smoothness assumption ensures that the point estimates are close to

the true graph divergence. The bounded moments assumption helps to handle the tails.

Definition 5.4.1. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a measurable set, $s > 0$, then C_L^β consists of all functions $g : \bar{\mathcal{X}} \to \mathbb{R}$ with continuous partial derivatives in $\mathcal{X}$ of integer order $s \leq \lfloor \beta \rfloor$, $D^s g$, such that*

$$\sum_{s=0}^{\lfloor \beta \rfloor} \sup_{x \in \bar{\mathcal{X}}} |D^s g(x)| + \sup_{x \neq y \in \bar{\mathcal{X}}} \frac{\left| D^{\lfloor \beta \rfloor} g(x) - D^{\lfloor \beta \rfloor} g(y) \right|}{\|x - y\|_\infty^{\beta - \lfloor \beta \rfloor}} \leq L \ . \tag{5.37}$$

Theorem 5.4.2. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a probability space and for $W \subseteq [d]$, let $p_W = \frac{dP_W}{d\lambda^d W}$ be the marginal density. Assume*

1. *For each $W \subseteq [d]$, $p_W \in C_L^\beta$ (Order β Hölder smooth)*
2. *$\mathbb{E}_P \left[\|X\|_\infty^\alpha \right] \leq C_b < \infty$*

If $X^{(1)}, \ldots, X^{(n)} \sim P$ and $k \in [n]$, then for all $\delta > 0$,

$$\left| \mathbb{E}\left[\widehat{D}_n(\mathcal{G}) \right] - D(\mathcal{G}) \right| \leq \mathcal{O}\left(n^{-\frac{\beta}{d+\beta} \frac{\alpha}{d+\alpha} + \delta} \right) \ . \tag{5.38}$$

The proof sketch can be found in appendix 5.8. As of now, this proof is not completely finished but most of the pieces are there. However, our goal is to more accurately calculate the bias so that we can remove it from the estimator to allow for a bias small enough to accommodate the central limit theorem.

Theorem 5.4.3. *Using the assumptions from theorem 5.4.2,*

$$Var[D_n(\mathcal{G})] \leq \frac{8d^2 \gamma_d}{n} \ . \tag{5.39}$$

Proof can be found in the appendix 5.8.

5.5 Mixed Graph Divergence

Many application fields of statistics, data science, and/or machine learning regularly have datasets where variables (or features) can be discrete or continuous. A probability measure is said to be continuous if it is absolutely continuous with respect to Lebesgue measure (λ) and discrete if its support is countable, and thus, is absolutely continuous with respect to counting measure over a countable set, C. A probability measure can be also be hybrid, discrete and continuous. This can happen if a random vector has some discrete coordinates and some continuous coordinates or if some coordinates themselves are mixed.

Definition 5.5.1. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional probability space such that $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{B} = \prod_{j=1}^{d} \mathcal{B}_j$ is the product σ-algebra, and P_j is the corresponding marginal probability measure. If for some $j \in [d]$, there exists a set $E \subseteq \mathcal{X}_j \subseteq \mathbb{R}$ such that $P_j(E) > 0$ and $P_j(\mathcal{X}\backslash E) > 0$ with $P_j \ll \lambda$ on E and $\mathcal{X}\backslash E$ countable then we say that P is a mixed probability measure.*

However, we place restrictions to ensure the existence of Radon-Nikodym (RN) derivatives.

Lemma 5.5.1. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional, mixed probability space. If P is non-singular then there exists a d-dimensional product measure μ on $\mathcal{X}$ such that $P \ll \mu$ such that for each $j \in [d]$, $\mu_j = \lambda + I_C$ where $C = \{x \in \mathcal{X}_j : P_j(\{x\}) > 0\}$.*

Definition 5.5.2. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional probability space. For each, $x \in \mathcal{X}$, let*

$$C^{(x)} = \left\{ i \in [d] : P_i^{(x)}(0) = 0 \right\} \tag{5.40}$$

be the indices of the continuous coordinates of x and

$$D^{(x)} = \left\{ i \in [d] : P_i^{(x)}(0) > 0 \right\} \tag{5.41}$$

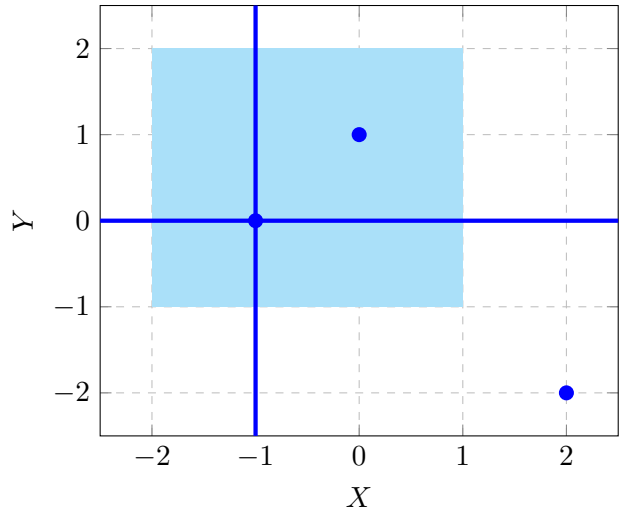


Figure 5.1: The figure above shows the support of a probability measure in $\mathbb{R}^2$. The cyan square indicates a region where both, X and Y are continuous. The dark blue lines indicate regions where X is continuous but Y is discrete and visa versa. The blue dots indicate where X and Y are both discrete.

the indices of the discrete coordinates of x. *Let* $x_D := \left(x_j : j \in D^{(x)}\right)$ *and* $x_C := \left(x_j : j \in C^{(x)}\right)$.

Lemma 5.5.2. *There exists a countable partition,* $\mathbf{T}$, *of* $\mathcal{X}$ *such that* $x, y \in A \in \mathbf{T}$ *implies that* $C^{(x)} = C^{(y)}$, $\frac{dP}{d\mu}$ *is uniformly (and absolutely) continuous on* A *and* $x_D = y_D$.

In general, this partitions the space into classes that are equal on discrete coordinates with continuous pdf for the continuous coordinates. Note that singletons satisfying these requirements are point masses in each dimension of $\mathcal{X}$.

Theorem 5.5.3. *Let* $(\mathcal{X}, \mathcal{B}, P)$ *be a non-singular* d*-dimensional, mixed probability space such that for each* $j \in [d]$, *if there exists a countable set* $E \subseteq \mathcal{X}_j$ *with* $P_j(E)$ *then* E *is nowhere dense* $[\lambda]$ *in* $\mathcal{X}_j$. *Let* $P_{\mathcal{G}}$ *be a Bayesian factorization with respect to DAG,* $\mathcal{G}$. *If* $\frac{dP}{dP_{\mathcal{G}}}$ *is the Radon-Nikodym derivative of* P *with respect to* $P_{\mathcal{G}}$, *then*

$$\frac{dP^{(x)}(r)}{dP_{\mathcal{G}}^{(x)}(r)} \to \frac{dP}{dP_{\mathcal{G}}}(x) \tag{5.42}$$

almost surely $[P]$ *on the support of* P*. Further, for the partition,* $\mathbf{T}$*, in lemma 5.5.2,*

$$D(\mathcal{G}) = \sum_{A\in\mathbf{T}} \int_A \log \frac{dP}{dP_G} dP \tag{5.43}$$

$$= \sum_{A\in\mathbf{T}} \left[\frac{P_D^{(x)}}{P_{G,D}^{(x)}} \log \left(\frac{P_D^{(x)}}{P_{G,D}^{(x)}} \right) I_A(x) + \int_A \log \left(\frac{dP_C}{dP_{G,C}}(x) \right) dP_C(x) \right] \tag{5.44}$$

where we define $0 \log 0 := 0$ *and recall that* $P_\varnothing = 1$*.*

Note that for the integral above, the discrete coordinates are necessary parameters because they provide location (conditioning) information. The continuous coordinates are not necessary for the discrete term but we leave them for notation consistency. Entropies are calculated in the standard (discrete and continuous) way though the probabilities may not add/integrate to one:

$$D(\mathcal{G}) = H(P) - \sum_{j=1}^{d} \left[H(P_{\mathrm{pa}^*(j)}) - H(P_{\mathrm{pa}(j)}) \right] \tag{5.45}$$

$$= \sum_{A\in\mathbf{T}} \left[P_D^{(x)} \log \left(P_D^{(x)} \right) I_A(x) + \int_A \log \left(\frac{dP_C}{d\lambda}(x) \right) dP_C(x) \right. \tag{5.46}$$

$$- \sum_{j=1}^{d} \left(P_{D\cap\mathrm{pa}^*(j)}^{(x)} \log \left(P_{D\cap\mathrm{pa}^*(j)}^{(x)} \right) I_A(x) + \int_A \log \left(\frac{dP_{C\cap\mathrm{pa}^*(j)}}{d\lambda}(x) \right) dP_{C\cap\mathrm{pa}^*(j)}(x) \right) \tag{5.47}$$

$$\left. + \sum_{j=1}^{d} \left(P_{D\cap\mathrm{pa}(j)}^{(x)} \log \left(P_{D\cap\mathrm{pa}(j)}^{(x)} \right) I_A(x) + \int_A \log \left(\frac{dP_{C\cap\mathrm{pa}(j)}}{d\lambda}(x) \right) dP_{C\cap\mathrm{pa}(j)}(x) \right) \right] \tag{5.48}$$

where the last three lines each correspond to an entropy estimate.

The previous chapter showed a method for estimating conditional mutual information for mixed data that can be directly applied to graph divergence as well.

Choose $k \in \mathbb{N} := \{1, 2, 3, \dots\}$ and let $W \subseteq [d] := \{1, 2, \dots, d\}$. Assume $(\mathcal{X}, \mathcal{B}, P)$ is a probability space and $X, X^{(1)}, X^{(2)}, \dots, X^{(n)} \sim P$ is an i.i.d. sample from P.

Consider the set of distances $\left\{\left\|X - X^{(i)}\right\|_\infty : i \in [n]\right\}$; let $\rho_k^{(X)}$ be the kth smallest element of this set, the kth nearest neighbor distance (if the reference point, X, is clear, we will drop the superscript for brevity). For $W \subseteq [d]$ where $X_W = (X_j : j \in W)$, define

$$n_{k,W}^{(X)} := \left|\left\{X^{(i)} : \left\|X_W - X_W^{(i)}\right\|_\infty \leq \rho_k^{(X)}\right\}\right|. \tag{5.49}$$

We may drop the k subscript when k is clear. This is the number of points within the kNN distance within a particular subspace of $\mathcal{X}$. Note that $n_{[d]}^{(X)}$ can be more than k if at least one of the coordinate in $[d]$ is discrete. For each $i \in [n]$, compute

$$\xi_k^{(i)} := \psi(k) - \psi(n+1) - \sum_{j=1}^{d}\left[\psi\left(n_{k,\mathrm{pa}^*(j)}^{(i)} + 1\right) - \psi\left(n_{k,\mathrm{pa}(j)}^{(i)} + 1\right)\right] \tag{5.50}$$

where $n_\varnothing^{(i)} = n$. Define

$$\widehat{D}_n(G) := \frac{1}{n}\sum_{i=1}^{n}\xi_k^{(i)}. \tag{5.51}$$

Theorem 5.5.4. *If $k = k_n \to \infty$, $\frac{k_n}{n} \to \infty$, $\frac{dP}{dP_\mathcal{G}} < \infty$ on $\mathcal{X}$, and the set $\{x \in \mathcal{X} : P_j(\{x\}) > 0, j \in [d]\}$ is nowhere dense in $\mathcal{X}$, then*

$$\mathbb{E}\left[\widehat{D}_n(\mathcal{G})\right] \to D(\mathcal{G}) \text{ and } Var\left[\widehat{D}_n(\mathcal{G})\right] \to 0 \tag{5.52}$$

as $n \to \infty$.

Proof. This proof follows directly from theorem 3.1 and 3.2 in the previous chapter. For conditional mutual information, $d = 3$, though this is a generalization of CMI, the same principles hold. □

5.6 Future Work: Greedy Equivalence Search

The previous sections showed how to calculate graph divergence given a particular Bayesian network. This section, in contrast, provides a heuristic for using graph divergence to estimate the Markov equivalence class (see [6, Sect. 3.4.3.3]) underlying a dataset. The heuristic we describe here is based on greedy equivalence search (GES) [4]. At a high level, GES takes a greedy approach to determine model fit, starting with an empty graph and adding one edge at a time to optimize a fit score. When the score can no longer be optimized by adding edges, the algorithm then begins removing edges. Reference [82] showed that an intermediate step of switching edge direction can improve performance. GES typically uses a composite Bayesian Information Criterion (BIC) to assess the fit of a Bayesian network to a dataset. Because BIC is biased toward sparser models, GES also tends to generate sparse Markov equivalence classes. Controlling the number of edges is necessary because, as explained in Section 5.2, it is possible to add edges to the true Bayesian network without changing score. The previous sections primarily focused on bounding asymptotic MSE. For this this reason, GES as intended for a composite BIC may not work as intended. This section is intended as a proof of concept.

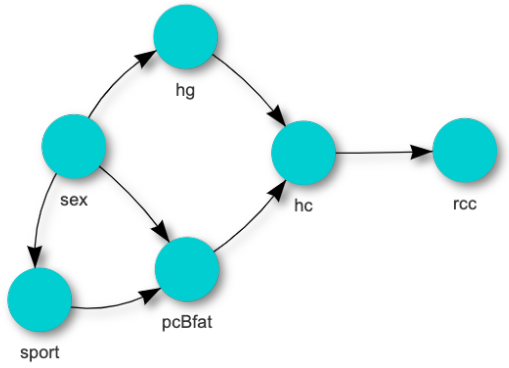
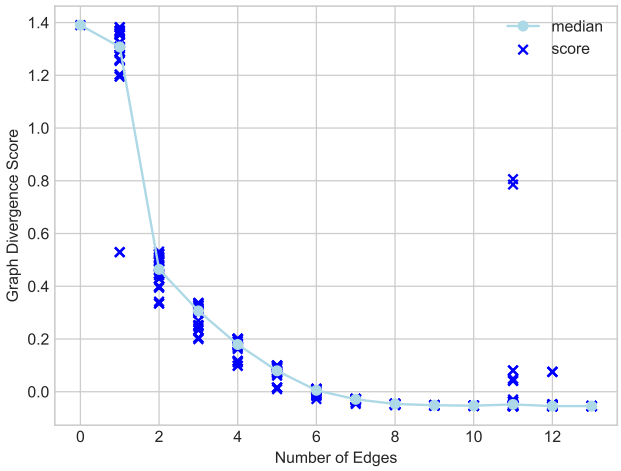
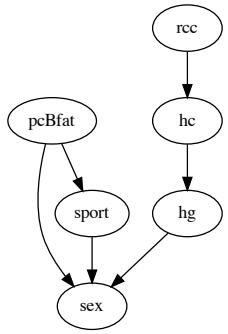


Figure 5.2: Simulated Bayesian network using the Australian Institute of Sport dataset from the R programming language DAAG package. Image copied from `https://www.r-bloggers.com/simulating-data-with-bayesian-networks/`

We simulated 1000 observations using the `bnlearn` package in R with the Bayesian network in figure 5.2 as in [83]. Combining the graph divergence score for mixed data and GIES [82], we collected all graph scores indicated by the algorithm. Figure 5.3a shows graph divergence scores for all graphs in the GIES search path by number of edges. Because this estimator is not biased toward sparser graphs, the final graph chosen will likely have too many edges.



(a) Scatter plot of graph divergence scores using GES



(b) Optimal Estimated DAG

Figure 5.3: Scatter plot of scores and estimated DAG on the Australian Institute of Sport simulation dataset.

Given these limitations, we used the elbow method as a heuristic to guide graph selection and selected the optimal six-edge graph, displayed in figure 5.3b. However, more testing is needed.

5.7 Conclusion

This work is unfinished. Currently, our plan is to consider making this into three papers: one for the continuous case, one for the mixed case, and one working out how best to use this estimator for graph divergence to search a graph space efficiently.

5.8 Proofs

Lemma 5.8.1 (5.4.1)**.** *If $X^{(1)}, \dots, X^{(n)} \sim P$ is a random sample (superscripts in parenthesis indicate observation number) and*

$$\xi^{(X)} := \psi(k) - \psi(n+1) - \sum_{j=1}^{d} \left[\psi\left(n_{pa^*(j)}^{(X)} + 1 \right) - \psi\left(n_{pa(j)}^{(X)} + 1 \right) \right], \tag{5.53}$$

then

$$\xi^{(X)} = \mathbb{E}\left[\log f_{\rho_k}(X) | X, n_{pa} \right] \tag{5.54}$$

where $n_{pa} := \left\{ n_{pa(j)}, n_{pa^*(j)} : j \in [d] \right\}$.

Proof of lemma 5.4.1. For any sample, $X, X^{(1)}, \dots, X^{(n)} \sim Q$,

$$Q^{(X)}\left(\left\| X - X^{(i)} \right\|_\infty \right) \sim \text{Unif}(0,1) \ . \tag{5.55}$$

Then for $k \in [n]$, kNN distance, $\rho_k^{(X)} = \left\| X_{(k)} - X \right\|_\infty$, with [84, Corollary 1.2],

$$Q^{(X)}(\rho_k^{(X)}) \sim \text{Beta}(k, n-k+1). \tag{5.56}$$

Moreover, for a discrete random variable, $K \in [n]$

$$\mathbb{E}\left[\log Q^{(X)}(\rho_K) \middle| K \right] = \psi(K) - \psi(n+1), \tag{5.57}$$

using the fact that $\mathbb{E}\left[\log\left(\text{Beta}(\alpha,\beta)\right)\right] = \psi(\alpha) - \psi(\alpha+\beta)$ where $\psi(x) := \frac{d}{dx}\log\Gamma(x)$ is the digamma function.

Note that $\rho_k^{(X)}$ is the kth order statistics of $\left\{ \left\| X - X^{(i)} \right\|_\infty : i \in [n] \right\}$. For a subsets of the dimension of X, $W \in [d]$,

$$n_W := \left| \left\{ i \in [n] : \left\| X_W - X_W^{(i)} \right\|_\infty < \rho_k^{(X)} \right\} \right| \tag{5.58}$$

is the number of sample points within the ℓ_∞ radius of X to its kNN in the dimen-

sions in W. For each point, let its ℓ_∞ distance to X be $r_{W,i}^{(X)} := \left\| X_W - X_W^{(i)} \right\|_\infty$ and $\rho_W^{(X)} = \max_{r_{W,i}^{(X)} \leq \rho_k^{(X)}} r_{W,i}^{(X)}$. Then

$$\log\left(P_W^{(X)}\left(\rho_k\right)\right) - \log\left(P_W^{(X)}\left(\rho_W\right)\right) = -\log\left(\frac{P_W^{(X)}\left(\max_{r_{W,i}\leq\rho_k} r_{W,i}\right)}{P_W^{(X)}\left(\rho_k\right)}\right) \tag{5.59}$$

$$= \min_{r_{W,i}\leq\rho_k} -\log\left(\frac{P_W^{(X)}\left(r_{W,i}\right)}{P_W^{(X)}\left(\rho_k\right)}\right) \tag{5.60}$$

$$= \min_{r_{W,i}\leq\rho_k} -\log\left(P_W\left(B_W^{(X)}\left(r_{W,i}\right)\middle| r_{W,i}\leq\rho_k\right)\right) \ . \tag{5.61}$$

Because $U_i := P_W\left(B_W^{(X)}\left(r_{W,i}\right)\middle| r_{W,i}\leq\rho_k\right) \sim \text{Unif}(0,1)$, then $-\log U_i \sim \text{exponential}(1)$ and

$$\min_{r_{W,i}\leq\rho_k} -\log\left(P_W\left(B_W^{(X)}\left(r_{W,i}\right)\middle| r_{W,i}\leq\rho_k\right)\right) \sim \text{exponential}(n_W) \tag{5.62}$$

as there are n_W points such that $r_{W,i} \leq \rho_k$. Then

$$\mathbb{E}\left[\log\left(P_W^{(X)}\left(\rho_k\right)\right) - \log\left(P_W^{(X)}\left(\rho_W\right)\right)\middle| n_W^{(X)}\right] = \frac{1}{n_W^{(X)}} \ . \tag{5.63}$$

Using the property of the digamma function, $\psi(x+1) = \psi(x) + \frac{1}{x}$,

$$\psi\left(n_W^{(X)}+1\right) - \psi(n+1) = \psi\left(n_W^{(X)}\right) - \psi(n+1) + \frac{1}{n_W^{(X)}} \tag{5.64}$$

$$= \mathbb{E}\left[\log P_W^{(X)}\left(\rho_W\right)\middle| n_W^{(X)}\right] + \mathbb{E}\left[\log\left(P_W^{(X)}\left(\rho_k\right)\right) - \log\left(P_W^{(X)}\left(\rho_W\right)\right)\middle| n_W^{(X)}\right] \tag{5.65}$$

$$= \mathbb{E}\left[\log\left(P_W^{(X)}\left(\rho_k\right)\right)\middle| n_W^{(X)}\right] \ . \tag{5.66}$$

Then

$$\xi^{(X)} = \psi(k) - \psi(n+1) - \sum_{j=1}^{d} \left[\psi\left(n^{(X)}_{\mathrm{pa}^*(j)} + 1\right) - \psi\left(n^{(X)}_{\mathrm{pa}(j)} + 1\right)\right] \tag{5.67}$$

$$= \psi(k) - \psi(n+1) - \sum_{j=1}^{d} \left[\psi\left(n^{(X)}_{\mathrm{pa}^*(j)} + 1\right) - \psi(n+1) - \psi\left(n^{(X)}_{\mathrm{pa}(j)} + 1\right) + \psi(n+1)\right] \tag{5.68}$$

$$= \mathbb{E}\left[\log\left(P^{(X)}(\rho_k)\right)\right] - \sum_{j=1}^{d} \mathbb{E}\left[\log\left(P^{(X)}_{\mathrm{pa}^*(j)}(\rho_k)\right)\middle| n^{(X)}_{\mathrm{pa}^*(j)}\right] - \mathbb{E}\left[\log\left(P^{(X)}_{\mathrm{pa}(j)}(\rho_k)\right)\middle| n^{(X)}_{\mathrm{pa}(j)}\right] \tag{5.69}$$

$$= \mathbb{E}\left[\log\left(\frac{P^{(X)}(\rho_k)\prod_{j=1}^{d} P^{(X)}_{\mathrm{pa}(j)}(\rho_k)}{\prod_{j=1}^{d} P^{(X)}_{\mathrm{pa}^*(j)}(\rho_k)}\right)\middle| n^{(X)}_{\mathrm{pa}}\right] \tag{5.70}$$

$$= \mathbb{E}\left[\log\left(\frac{P^{(X)}(\rho_k)}{P^{(X)}_{\mathcal{G}}(\rho_k)}\right)\middle| n^{(X)}_{\mathrm{pa}}\right] = \mathbb{E}\left[\log f_{\rho_k}(X)\middle| n^{(X)}_{\mathrm{pa}}\right] \ . \tag{5.71}$$

□

Lemma 5.8.2. *Assume $X \sim P$ with density, p, and $\mathbb{E}\left[\|X\|_\infty^\alpha\right] < \infty$. Then for $\tau \in \left(0, \frac{\alpha}{\alpha+d}\right)$, and constant, C,*

$$\mathbb{P}\left(p(X) < t\right) \leq Ct^\tau \tag{5.72}$$

and

$$\int_{\mathcal{X}} [p(x)]^s \exp\left\{-bp(x)\right\} dx \leq \frac{K_s}{b^{s+\tau-1}} \tag{5.73}$$

Proof. Using Hölder's inequality,

$$\mathbb{E}\left[p^{-\tau}(X)\right] = \int_{\mathcal{X}} p^{1-\tau}(x)dx \tag{5.74}$$

$$= \int_{\mathcal{X}} \left[\left(1+\|x\|^{\alpha}\right)p(x)\right]^{1-\tau}\left(\frac{1}{1+\|x\|^{\alpha}}\right)^{1-\tau} dx \tag{5.75}$$

$$\leq \left(\int_{\mathcal{X}} \left(1+\|x\|^{\alpha}\right)p(x)dx\right)^{1-\tau}\left(\int_{\mathcal{X}}\left(\frac{1}{1+\|x\|^{\alpha}}\right)^{\frac{1-\tau}{\tau}} dx\right)^{\tau} \tag{5.76}$$

$$= \left(1+\mathbb{E}\left[\|X\|^{\alpha}\right]\right)^{1-\tau}\left(\int_{\mathcal{X}}\left(\frac{1}{1+\|x\|^{\alpha}}\right)^{\frac{1-\tau}{\tau}} dx\right)^{\tau} := C\ . \tag{5.77}$$

The second factor of the last line is finite if $\tau < \frac{\alpha}{\alpha+d}$.

Let $\tau < \frac{\alpha}{\alpha+d}$. Then

$$\mathbb{P}(p(X) < t) = \mathbb{P}\left(p^{-\tau}(X) > t^{-\tau}\right) \tag{5.78}$$

$$\leq t^{\tau}\mathbb{E}\left[p^{-\tau}(X)\right] \qquad \text{Markov's Inequality} \tag{5.79}$$

$$= Ct^{\tau}\ . \tag{5.80}$$

$$\int_{\mathcal{X}} [p(x)]^s \exp\{-bp(x)\}\, dx = \mathbb{E}\left[[p(x)]^{s-1}\exp\{-bp(x)\}\right] \tag{5.81}$$

$$= \frac{1}{b^{s-1}}\mathbb{E}\left[[bp(x)]^{s-1}\exp\{-bp(x)\}\right] \tag{5.82}$$

$$\leq \frac{2(s-1)e^{-1}}{b^{s-1}}\mathbb{E}\left[\exp\left\{\frac{b}{2}p(x)\right\}\exp\{-bp(x)\}\right] \tag{5.83}$$

$$\leq \frac{K_s}{b^s} \tag{5.84}$$

□

Lemma 5.8.3. *Assume that* $\mathbb{P}(p(X) \leq t) \leq Ct^{\tau}$ *for* $\tau \in (0,1)$*, then for a constant,* C*,*

$$\lambda\{x \in \mathcal{X} : p(X) \geq t\} \leq \frac{\tau}{1-\tau}Ct^{\tau-1}. \tag{5.85}$$

Proof. See [81, Lemmas 3 and Theorem 6]. □

Lemma 5.8.4. *Let $p_W \in C_L^\beta$, $W \subseteq [d]$, and $d_W := |W|$, then for any $x \in \mathcal{X}$,*

$$\left| P_W^{(x)}(r) - p_W(x) r^{d_W} \right| \leq L r^{\beta + d_W} \ . \tag{5.86}$$

Proof. We use a $\lfloor \beta \rfloor$th Taylor expansion of $p_W(y)$ about x, with the intermediate value theorem for the remainder,

$$|p_W(y) - p_W(x)| = \left| \sum_{t=0}^{\lfloor \beta \rfloor - 1} \frac{\|y-x\|^t}{t!} D^t p_W(x) + \frac{\|y-x\|^{\lfloor \beta \rfloor}}{\lfloor \beta \rfloor !} D^{\lfloor \beta \rfloor} p_W(\zeta) - p_W(x) \right| \tag{5.87}$$

$$= \left| \sum_{t=1}^{\lfloor \beta \rfloor} \frac{\|y-x\|^t}{t!} D^t p_W(x) + \frac{\|y-x\|^{\lfloor \beta \rfloor}}{\lfloor \beta \rfloor !} \left(D^{\lfloor \beta \rfloor} p_W(\zeta) - D^{\lfloor \beta \rfloor} p_W(x) \right) \right| \tag{5.88}$$

$$= \|y-x\|^\beta \left| \sum_{t=1}^{\lfloor \beta \rfloor} \frac{\|y-x\|^{t - \lfloor \beta \rfloor}}{t!} D^t p_W(x) + \frac{D^{\lfloor \beta \rfloor} p_W(\zeta) - D^{\lfloor \beta \rfloor} p_W(x)}{\lfloor \beta \rfloor ! \, \|y-x\|^{\beta - \lfloor \beta \rfloor}} \right| \tag{5.89}$$

$$\leq \|y-x\|^\beta \left| \sum_{t=1}^{\beta} D^t p_W(x) + \frac{D^{\lfloor \beta \rfloor} p_W(\zeta) - D^{\lfloor \beta \rfloor} p_W(x)}{\|y-x\|^{\beta - \lfloor \beta \rfloor}} \right| \tag{5.90}$$

$$\leq L \|y-x\|^\beta \ . \tag{5.91}$$

The last line follow because $p_W \in C_L^\beta$. For $r > 0$, we have (where the integral is

with respect to the d_W-dimensional Lebesgue measure)

$$\left|P_W^{(x)}(r) - p_W(x) r^{d_W}\right| = \left|\int_{B^{(x)}(r)} p_W(y)dy - p_W(x)\right| \tag{5.92}$$

$$= \left|\int_{B^{(x)}(r)} p_W(y) - p_W(x)dy\right| \tag{5.93}$$

$$\leq \int_{B^{(x)}(r)} |p_W(y) - p_W(x)|\, dy \tag{5.94}$$

$$\leq \int_{B^{(x)}(r)} Lr^{\beta} dy = Lr^{\beta+d_W}\ . \tag{5.95}$$

□

Lemma 5.8.5. *Assume that $g \in C_L^{\beta}(\mathbb{R}^d)$ be a density and $G^{(x)}(r) := \int_{B^{(x)}} g(y)dy$ (integral with respect to the Lebesgue measure). If $W \subseteq [d]$ then*

$$\left|G_W^{(x)}(r) - g_W(x) r^{d_W}\right| \leq C\left[G^{(x)}(r)\right]^{\beta/d} \tag{5.96}$$

for a constant, C, with respect to r.

Proof. TBD □

proof of Theorem 5.4.2. Let $X \sim P$ and $S = \{x \in \mathcal{X} : p(x) \leq Ln^{-\gamma}\}$ for $\gamma > 0$.

Then

$$\mathbb{E}\left[\widehat{D}_n(\mathcal{G})\right] - D(\mathcal{G}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\xi^{(i)}\right] - \mathbb{E}\left[\log\left(\frac{dP}{dP_{\mathcal{G}}}(X)\right)\right] \tag{5.97}$$

$$= \mathbb{E}\left[\xi^{(X)}\right] - \mathbb{E}\left[\log\left(\frac{dP}{dP_{\mathcal{G}}}(X)\right)\right] \tag{5.98}$$

$$= \mathbb{E}\left[\left(\xi^{(X)} - \log\left(\frac{dP}{dP_{\mathcal{G}}}(X)\right)\right) I(X \in S)\right] \tag{5.99}$$

$$+ \mathbb{E}\left[\left(\xi^{(X)} - \log\left(\frac{dP}{dP_{\mathcal{G}}}(X)\right)\right) I(X \notin S)\right] \tag{5.100}$$

$$= \mathbb{E}\left[\xi^{(X)} I(X \in S)\right] - \mathbb{E}\left[\log\left(\frac{dP}{dP_{\mathcal{G}}}(X)\right) I(X \in S)\right] \tag{5.101}$$

$$+ \mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{P_{\mathcal{G}}^{(X)}(\rho_k)}\right) - \log\left(\frac{dP}{dP_{\mathcal{G}}}(X)\right)\right) I(X \notin S)\right] \tag{5.102}$$

where line 5.102 follows from lemma 5.4.1.

We proceed by analyzing lines 5.101 and 5.102 separately. We begin with the second term of line 5.101. Let $T := p(X)$ and $F_T(t) := \mathbb{P}(T \leq t) \leq Ct^{\tau}$, where the inequality comes from lemma 5.8.2. Then

$$\mathbb{E}\left[\log p(X) I(X \in S)\right] = \mathbb{E}\left[\log T I(X \in S)\right] \tag{5.103}$$

$$= \mathbb{E}\left[\log T I\left(T < Ln^{-\gamma}\right)\right] \tag{5.104}$$

$$= \int_0^{Ln^{-\gamma}} \log t \, dF_T(t) \tag{5.105}$$

$$= \log(t) F_T(t)\big|_0^{Ln^{-\gamma}} - \int_0^{Ln^{-\gamma}} \frac{F_T(t)}{t} dt \tag{5.106}$$

$$\geq \log\left(Ln^{-\gamma}\right)\left(C(Ln^{-\gamma})^{\tau}\right) - \int_0^{Ln^{-\gamma}} \frac{Ct^{\tau}}{t} dt \tag{5.107}$$

$$= \log\left(Ln^{-\gamma}\right)\left(C(Ln^{-\gamma})^{\tau}\right) - \frac{C\left(Ln^{-\gamma}\right)^{\tau}}{\tau} \tag{5.108}$$

$$= -\mathcal{O}\left(n^{-\gamma\tau}\log n\right) \ . \tag{5.109}$$

Using this,

$$\left|\mathbb{E}\left[\log\left(\frac{dP}{dP_{\mathcal{G}}}(X)\right) I(X \in S)\right]\right| \tag{5.110}$$

$$= \left|\mathbb{E}\left[\left(\log p(X) - \sum_{j=1}^{d} \log p_{\mathrm{pa}^*(j)}(X) - \log p_{\mathrm{pa}(j)}(X)\right) I(X \in S)\right]\right| \tag{5.111}$$

$$\leq \left|\mathbb{E}\left[\left(\log p(X) + \sum_{j=1}^{d} \log p_{\mathrm{pa}(j)}(X)\right) I(X \in S)\right]\right| \tag{5.112}$$

$$\leq (d+1)\left|\mathbb{E}\left[\log p(X) I(X \in S)\right]\right| \qquad -\log p_W(x) \leq -\log p(x) \tag{5.113}$$

$$= \mathcal{O}\left(n^{-\gamma\tau}\log n\right) \ . \tag{5.114}$$

Moving to the first term of line 5.101,

$$\left|\mathbb{E}\left[\xi^{(X)} I(X \in S)\right]\right| \tag{5.115}$$

$$= \left|\mathbb{E}\left[\left(\psi(k) - \psi(n+1) - \sum_{j=1}^{d} \psi\left(n^{(X)}_{\mathrm{pa}^*(j)} + 1\right) - \psi\left(n^{(X)}_{\mathrm{pa}(j)} + 1\right)\right) I(X \in S)\right]\right| \tag{5.116}$$

$$\leq \left|\mathbb{E}\left[\left(-\psi(n+1) - \sum_{j=1}^{d} \psi\left(n^{(X)}_{\mathrm{pa}^*(j)} + 1\right)\right) I(X \in S)\right]\right| \tag{5.117}$$

$$\leq (d+1)\left|\mathbb{E}\left[\psi(n+1) I(X \in S)\right]\right| \qquad n^{(X)}_W \leq n \tag{5.118}$$

$$\leq (d+1)\log(n+1)\mathbb{P}(X \in S) \qquad \psi(x) \leq \log(x) \tag{5.119}$$

$$\leq (d+1) L n^{-\gamma}\log n \ . \tag{5.120}$$

Together, these show that line 5.101 is bounded by $\mathcal{O}\left(n^{-\gamma\tau}\log n\right)$.

Moving to line 5.102,

$$\left|\mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{P_{\mathcal{G}}^{(X)}(\rho_k)}\right)-\log\left(\frac{dP}{dP_{\mathcal{G}}}(X)\right)\right)I(X\notin S)\right]\right| \tag{5.121}$$

$$=\left|\mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{P_{\mathcal{G}}^{(X)}(\rho_k)}\right)-\log\left(\frac{p(x)\rho_k^d}{p_{\mathcal{G}}(x)\rho_k^d}\right)\right)I(X\notin S)\right]\right| \tag{5.122}$$

$$=\left|\mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{p(x)\rho_k^d}\right)-\log\left(\frac{P_{\mathcal{G}}^{(X)}(\rho_k)}{p_{\mathcal{G}}(x)\rho_k^d}\right)\right)I(X\notin S)\right]\right| \tag{5.123}$$

$$=\left|\mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{p(x)\rho_k^d}\right)-\sum_{j=1}^{d}\log\left(\frac{P_{\mathrm{pa}^*(j)}^{(X)}(\rho_k)}{p_{\mathrm{pa}^*(j)}(X)\rho_k^{d_{\mathrm{pa}^*(j)}}}\right)-\log\left(\frac{P_{\mathrm{pa}(j)}^{(X)}(\rho_k)}{p_{\mathrm{pa}(j)}(X)\rho_k^{d_{\mathrm{pa}(j)}}}\right)\right)I(X\notin S)\right]\right| \tag{5.124}$$

$$\leq\left|\mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{p(x)\rho_k^d}\right)+\sum_{j=1}^{d}\log\left(\frac{P_{\mathrm{pa}(j)}^{(X)}(\rho_k)}{p_{\mathrm{pa}(j)}(X)\rho_k^{d_{\mathrm{pa}(j)}}}\right)\right)I(X\notin S)\right]\right| \tag{5.125}$$

$$\leq\left|\mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{p(x)\rho_k^d}\right)+\sum_{j=1}^{d}\log\left(\frac{P_{\mathrm{pa}(j)}^{(X)}(\rho_k)}{p_{\mathrm{pa}(j)}(X)\rho_k^{d_{\mathrm{pa}(j)}}}\right)\right)I(X\notin S,\rho_k>a_n)\right]\right| \tag{5.126}$$

$$+\left|\mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{p(x)\rho_k^d}\right)+\sum_{j=1}^{d}\log\left(\frac{P_{\mathrm{pa}(j)}^{(X)}(\rho_k)}{p_{\mathrm{pa}(j)}(X)\rho_k^{d_{\mathrm{pa}(j)}}}\right)\right)I(X\notin S,\rho_k\leq a_n)\right]\right| \tag{5.127}$$

$$\tag{5.128}$$

Moving to line 5.127, we use lemma 5.8.4:

$$\left|\mathbb{E}\left[\left(\log P_W^{(X)}(\rho_k) - \log\left(p_W(X)\rho_k^{d_W}\right)\right) I(X \notin S, \rho_k \leq a_n)\right]\right| \tag{5.129}$$

$$\leq \mathbb{E}\left[\left|\log P_W^{(X)}(\rho_k) - \log\left(p_W(X)\rho_k^{d_W}\right)\right| I(X \notin S, \rho_k \leq a_n)\right] \tag{5.130}$$

$$\leq \mathbb{E}\left[\max\left\{\left|\log\left(p_W(X)\rho_k^{d_W} \pm L\rho_k^{d_W+\beta}\right) - \log\left(p(X)\rho_k^{d_W}\right)\right|\right\} I(X \notin S, \rho_k \leq a_n)\right] \tag{5.131}$$

$$= \mathbb{E}\left[\left|\log\left(p_W(X)\rho_k - L\rho_k^{\beta}\right) - \log\left(p_W(X)\rho_k\right)\right| I(X \notin S, \rho_k \leq a_n)\right] \tag{5.132}$$

$$\leq \mathbb{E}\left[\frac{L\rho_k^{\beta}}{\zeta p_W(X)} I(X \notin S, \rho_k \leq a_n)\right] \tag{5.133}$$

$$\leq \mathbb{E}\left[\frac{L\rho_k^{\beta}}{\zeta p(X)} I(X \notin S, \rho_k \leq a_n)\right] \tag{5.134}$$

$$\leq \mathbb{E}\left[\frac{2L\rho_k^{\beta}}{p(X)} I(X \notin S, \rho_k \leq a_n)\right] \tag{5.135}$$

$$\leq \lambda\left\{\mathcal{X}\backslash S\right\} \tag{5.136}$$

$$\leq \frac{\tau}{1-\tau} C\left(n^{-\gamma}\right)^{\tau-1} \tag{5.137}$$

We use the intermediate value theorem on line 5.133, so that $1 - \frac{L\rho_k^{\beta}}{p_W(x)} \leq \zeta \leq 1$. The next line follows because $p(X) \leq p_W(X)$. For the same reason, we also have we have $1 - \frac{L\rho_k^{\beta}}{p(x)} \leq \zeta \leq 1$. And, with choice of S and a_n,

$$\frac{Lr^{\beta}}{p(x)} \leq \frac{La_n^{\beta}}{p(x)} \leq \frac{1}{2}. \tag{5.138}$$

So, $\zeta \geq \frac{1}{2}$. The last step uses lemma 5.8.3 where $\tau \in \left(0, \frac{\alpha}{\alpha+d}\right)$. Putting this together,

$$\left|\mathbb{E}\left[\left(\log\left(\frac{P^{(X)}(\rho_k)}{p(x)\rho_k^d}\right) + \sum_{j=1}^{d} \log\left(\frac{P_{\mathrm{pa}(j)}^{(X)}(\rho_k)}{p_{\mathrm{pa}(j)}(X)\rho_k^{d_{\mathrm{pa}(j)}}}\right)\right) I(X \notin S, \rho_k > a_n)\right]\right| \leq \frac{(d+1)\tau}{1-\tau} C n^{-\gamma(\tau-1)}. \tag{5.139}$$

For line 5.126, let $W \subseteq [d]$,

$$\left|\mathbb{E}\left[\left(\log P_W^{(X)}(\rho_k) - \log\left(p_W(x)\rho_k^d\right)\right) I(X \notin S, \rho_k > a_n)\right]\right| \tag{5.140}$$

$$\leq \left|\mathbb{E}\left[\log P_W^{(X)}(\rho_k) I(X \notin S, \rho_k > a_n)\right]\right| \tag{5.141}$$

$$\leq \left|\mathbb{E}\left[\log P_W^{(X)}(\rho_k) I(X \notin S, \rho_k > a_n)\right]\right| \tag{5.142}$$

$$\leq \left|\mathbb{E}\left[\log P_W^{(X)}(\rho_k) \middle| \rho_k > a_n\right]\right| \mathbb{P}(X \notin S, \rho_k > a_n) \tag{5.143}$$

$$\leq \left|\psi\left(n_{W^*}^{(X)}\right) - \psi\left(n_W^{(X)}\right)\right| \mathbb{P}(X \notin S, \rho_k > a_n) \tag{5.144}$$

$$\leq n^{-\gamma} \log n \tag{5.145}$$

Using [84, thm 20.3.1],

$$\mathbb{P}\left(\rho_k > a_n\right) = \mathbb{P}\left(\text{Binomial}\left(n, P^{(x)}\left(a_n\right)\right) \leq k\right) \tag{5.146}$$

$$\leq \exp\left\{k - nP^{(x)}(a_n) - k\log\left(\frac{k}{nP^{(x)}(a_n)}\right)\right\} \tag{5.147}$$

$$= \exp\left\{-nP^{(x)}(a_n)\right\}\left(\frac{enP^{(x)}(a_n)}{k}\right)^k. \tag{5.148}$$

Because $P^{(x)}(r) \geq p(x)r^d - r^{d+\beta} \geq \frac{1}{2}p(x)r^d$ and $\exp(-nx)\left(\frac{enx}{k}\right)^k$ is decreasing as a function of x,

$$\mathbb{P}(X \notin S, \rho_k > a_n) \leq \int_{\mathcal{X}\setminus S} \exp\left\{-nP^{(x)}(a_n)\right\}\left(\frac{enP^{(x)}(a_n)}{k}\right)^k dP(x) \tag{5.149}$$

$$\leq \int_{\{p(x)\geq Ln^{-\gamma}\}} \exp\left\{-n\left(\frac{1}{2}p(x)a_n^d\right)\right\}\left(\frac{en\left(\frac{1}{2}p(x)a_n^d\right)}{k}\right)^k dP(x) \tag{5.150}$$

$$\leq \int \exp\left\{-n\left(\frac{1}{2}Ln^{-\gamma}n^{-d}\right)\right\}\left(\frac{en\left(\frac{1}{2}Ln^{-\gamma}n^{-d}\right)}{k}\right)^k dP(x) \tag{5.151}$$

$$\leq Cn^{-(\beta+d)} \tag{5.152}$$

□

Proof of Theorem 5.4.3. Using Stein-Efron, let $x^{(1)}, \ldots, x^{(n)} \sim P$ and $y^{(1)}, \ldots, y^{(n)} \sim P$ be two independent samples. Let $Z = (x^{(1)}, \ldots, x^{(n)})$ and $Z^l = (x^{(1)}, \ldots, x^{(l-1)}, y^{(l)}, x^{(l+1)}, \ldots, x^{(n)})$.

$$\operatorname{Var}\left(\widehat{D}_n\right) \tag{5.153}$$

$$\leq \frac{1}{2}\sum_{l=1}^{n}\mathbb{E}\left[\widehat{D}_n(Z) - \widehat{D}_n(Z^l)\right]^2 \tag{5.154}$$

$$= \frac{1}{2}\sum_{l=1}^{n}\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\xi^{(i)}(Z) - \frac{1}{n}\sum_{i=1}^{n}\xi^{(i)}(Z^l)\right)^2\right] \tag{5.155}$$

$$= \frac{1}{2n^2}\sum_{l=1}^{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\left(\xi^{(i)}(Z) - \xi^{(i)}(Z^l)\right)\right)^2\right] \tag{5.156}$$

$$= \frac{1}{2n^2}\sum_{l=1}^{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{d}\left[\psi\left(n^{(i)}_{\mathrm{pa}^*(j)}(Z^l)\right) - \psi\left(n^{(i)}_{\mathrm{pa}^*(j)}(Z)\right)\right] + \left[\psi\left(n^{(i)}_{\mathrm{pa}(j)}(Z)\right) - \psi\left(n^{(i)}_{\mathrm{pa}(j)}(Z^l)\right)\right]\right)^2\right] \tag{5.157}$$

$$\leq \frac{1}{2n^2}\sum_{l=1}^{n}\mathbb{E}\left[\left(\sum_{j=1}^{d}\frac{\gamma_d n_{\mathrm{pa}^*(j)}}{n_{\mathrm{pa}^*(j)} - 1} + \frac{\gamma_d n_{\mathrm{pa}(j)}}{n_{\mathrm{pa}(j)} - 1}\right)^2\right] \tag{5.158}$$

$$\leq \frac{1}{2n^2}\sum_{l=1}^{n}\mathbb{E}\left[\left(\sum_{j=1}^{d}2\gamma_d + 2\gamma_d\right)^2\right] \tag{5.159}$$

$$= \frac{1}{2n^2}\sum_{l=1}^{n}\mathbb{E}\left[16d^2\gamma_d\right] = \frac{8d^2\gamma_d}{n} \tag{5.160}$$

Using Gyorfi, corollary 6.1,

$$\sum_{i=i}^{n} I_{X\in B^i_W\left(\rho_{n_W}\right)}\left(X^{(i)}\right) \leq n_W\gamma_{d_W}. \tag{5.161}$$

□

Let λ be the Lebesgue measure of dimension 1 unless otherwise specified and $[d] = \{1, \ldots, d\}$. Further, assume that all conditional probability measures are

regular.

Lemma 5.8.6. *5.5.1 Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional probability space such that for each $i \in [d]$, there exists a $\mathcal{B}_i$-measurable set $E \subseteq \mathcal{X}_i$ such that $P_i \ll \lambda$ on E and $\lambda(\mathcal{X} \backslash E) = 0$. If P is non-singular then there exists a d-dimensional product measure μ on $\mathcal{X}$ such that $P \ll \mu$ such that for each $i \in [d]$, $\mu_i = \lambda + I_C$ where $C = \{x \in \mathcal{X}_i : P_i(\{x\}) > 0\}$.*

Proof. We construct μ by looking at the scalar coordinates of the random variable $V \sim P$ over its product space, $\mathcal{X} \equiv \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$. If $\mathcal{X}_i$ is not a subset of $\mathbb{R}$, V_i is categorical and we use a zero-one distance metric. So that we can work exclusively in $\mathbb{R}^c$ for some positive integer c, we create dummy indicators for all categories except one; this preserves the ℓ_∞ metric for categorical variables. Recall that the marginal measure for any scalar coordinate is $P_{V_i}(A) = P_V(\mathcal{X}_1 \cdots \times \mathcal{X}_{i-1} \times A \times \mathcal{X}_{i+1} \times \cdots \times \mathcal{X}_d)$ where $A \subseteq \mathcal{X}_i$. For each $i = 1, \ldots, d$, redefine $\mathcal{X}_i$ by restricting it to the support of P_{V_i} and $\mathcal{B}_{\mathcal{X}_i}$ the corresponding σ-algebra. Partition $\mathcal{X}_i$ into its discrete and continuous parts. For a set A contained within the support of a random variable, U, let $C_U(A) = \{x \in A : P_U(x) = 0\}$ be the continuous partition and $D_U(A) = \{x \in A : P_U(x) > 0\}$, which is countable by assumption. Clearly $C_U(A) \cup D_U(A) = A$ and $C_U(A) \cap D_U(A) = \varnothing$ for all random variables U. Let λ be the Lebesgue measure and ν be the counting measure. Define the measure $\mu_i : \mathcal{B}_{\mathcal{X}_i} \to [0, \infty)$ to be $\lambda + \nu_i$, where $\nu_i(C_{V_i}(\mathcal{X}_i)) = 0$ and the counting measure on $D_{V_i}(\mathcal{X}_i)$, $\nu_i(D_{V_i}(\mathcal{X}_i)) = \nu(D_{V_i}(\mathcal{X}_i))$. It is easy to see that μ_i is a well-defined measure on the measurable space, $(\mathcal{X}_i, \mathcal{B}_{\mathcal{X}_i})$ because both the counting measure and Lebesgue measures are well-defined, as is their sum. Define a measure $\mu : \mathcal{B}_{\mathcal{X}} \to \mathbb{R}$ as the product measure, $\mu = \mu_1 \times \mu_2 \times \cdots \times \mu_d$.

With the construction complete, we now show that $P_V \ll \mu$. We begin by showing that for each coordinate, $j = 1, \ldots, d$, $P_{\mathcal{X}_j} \ll \mu_j$. Let $j = 1, \ldots, d$ and $A \in \mathcal{B}_{\mathcal{X}_i}$ with $\mu_j(A) = 0$. Consider the continuous and discrete partitions, $C_{V_j}(A)$ and $D_{V_j}(A)$, respectively. By definition, $\lambda(C_{V_j}(A)) + \nu_j(D_{V_j}(A)) = 0$ so $\lambda(C_{V_j}(A)) = 0$

and $\nu_j(D_{V_j}(A)) = 0$. If the coordinate project for j has a nonempty continuous partition, then $P_{V_j} \ll \lambda$ on $C_{V_j}(\mathcal{X}_j)$, so $P_{V_j}(C_{V_j}(A)) = 0$. Also, $0 = \nu_j(D_{V_j}(A)) = \nu(D_{V_j}(A))$, so $D_{V_j}(A) = \varnothing$, so $P_{V_j}(D_{V_J}(A)) = 0$. Then $P_{V_j}(A) = P_{V_j}(C_{V_j}(A)) + P_{V_j}(D_{V_j}(A)) = 0$

Proceeding by mathematical induction, we already have $P_{V_1} \ll \mu_1$. Assume that $P_{V_1 \dots V_j} \ll \mu_1 \times \cdots \times \mu_j \equiv \prod_{i=1}^{j} \mu_i$ and that for some $A \in \mathcal{B}_{\mathcal{X}_1 \dots \mathcal{X}_j \mathcal{X}_{j+1}}$ (the product σ-algebra) $(\prod_{i=1}^{j+1} \mu_i)(A) = 0$. Let $A_{v_1,\dots,v_j} = \{v_{j+1} : (v_1, \dots, v_j, v_{j+1}) \in A\}$ and $A_{v_{j+1}} = \{(v_1, \dots, v_j) : (v_1, \dots, v_j, v_{j+1}) \in A\}$. Let $A_1 = \mathcal{X}_1 \times \cdots \times \mathcal{X}_j \left\{v_{j+1} : P_{V_1 \dots V_j}(A_{v_{j+1}}) > 0\right\}$ and $A_2 = \left\{(v_1, \dots, v_j) : P_{V_{j+1}}(A_{v_1,\dots,v_j}) > 0\right\} \times \mathcal{X}_{j+1}$.

Using Fubini's theorem,

$$\begin{aligned} 0 &= \left(\prod_{i=1}^{j+1} \mu_i\right)(A) \\ &= \left(\prod_{i=1}^{j} \mu_i \times \mu_{j+1}\right)(A) \\ &= \int_{\mathcal{X}_{j+1}} \left(\prod_{i=1}^{j} \mu_i\right)(A_{v_{j+1}}) d\mu_{j+1}(v_{j+1}). \end{aligned}$$

Using Lemma 1.3.8 (dembo2019probability), $f \geq 0, \int f d\mu = 0 \Rightarrow \mu\{x : f(x) > 0\} = 0$, we must have

$$\begin{aligned} 0 &= \mu_{j+1}\left(\left\{v_{j+1} : \left(\prod_{i=1}^{j} \mu_i\right) A_{v_{j+1}}) > 0\right\}\right) \\ &= \mu_{j+1}\left(\mathcal{X}_{j+1} \backslash \left\{v_{j+1} : \left(\prod_{i=1}^{j} \mu_i\right)(A_{v_{j+1}}) = 0\right\}\right) \\ &\geq \mu_{j+1}\left(\mathcal{X}_{j+1} \backslash \left\{v_{j+1} : P_{V_1 \dots V_j}(A_{v_{j+1}}) = 0\right\}\right). \end{aligned}$$

The last inequality follows because $P_{V_1 \dots V_j} \ll \prod_{i=1}^{j} \mu_i$ implies that $\left\{v_{j+1} : \left(\prod_{i=1}^{j} \mu_i\right)(A_{v_{j+1}}) = 0\right\} \subseteq \left\{v_{j+1} : P_{V_1 \dots V_j}(A_{v_{j+1}}) = 0\right\}$. Then $\mu_{j+1}\left(\mathcal{X}_{j+1} \backslash \left\{v_{j+1} : P_{V_1 \dots V_j}(A_{v_{j+1}}) = 0\right\}\right) = 0$.

But, $P_{V_{j+1}} \ll \mu_{j+1}$ implies that

$$\begin{aligned} 0 &= P_{V_{j+1}}\left(\mathcal{X}_{j+1} \backslash \left\{v_{j+1} : P_{V_1 \ldots V_j}(A_{v_{j+1}}) = 0\right\}\right) \\ &= P_{V_{j+1}}\left(\left\{v_{j+1} : P_{V_1 \ldots V_j}(A_{v_{j+1}}) > 0\right\}\right) \\ &= P_{V_1 \ldots V_j V_{j+1}}(A_1). \end{aligned}$$

Using the same procedure but switching $\prod_{i=1}^{j} \mu_i$ and μ_{j+1} and correspondingly, switching $P_{V_1 \ldots V_j}$ and $P_{V_{j+1}}$, it is easy to show that

$$\begin{aligned} 0 &= P_{V_1 \ldots V_j}\left(\left\{(v_1, \ldots, v_j) : P_{V_{j+1}}(A_{v_1,\ldots,v_j}) > 0\right\}\right) \\ &= P_{V_1 \ldots V_j V_{j+1}}(A_2). \end{aligned}$$

Consider the set of points $(v_1, \ldots, v_j, v_{j+1})$ such that each coordinate satisfies $P_{V_{j+1}}(A_{v_1,\ldots,v_j}) = 0$ and $P_{V_1 \ldots V_j}(A_{v_{j+1}}) = 0$; call this set, A_3. Showing that $P(A_3) = 0$, consider the set of points, $(a_1, \ldots a_d) \in B \subseteq \mathcal{X}$ such that

$$P_{V_{j+1} \ldots V_d}\left(\left[A \times \prod_{i=j+2}^{d} \mathcal{X}_i\right]_{(a_1,\ldots,a_j)}\right) = 0$$

and

$$P_{V_1 \ldots V_j}\left(\left[A \times \prod_{i=j+2}^{d} \mathcal{X}_i\right]_{(a_{j+1},\ldots,a_d)}\right) = 0.$$

Let $(b_1, \ldots, b_d) \in A_3 \times \prod_{i=j+2}^{d} \mathcal{X}_i$. Then

$$\begin{aligned} & P_{V_{j+1} \ldots V_d} \left(\left[A \times \prod_{i=j+2}^{d} \mathcal{X}_i \right]_{(b_1, \ldots, b_j)} \right) \\ & = P_{V_{j+1} \ldots V_d} \left(A_{(b_1, \ldots, b_j)} \times \prod_{i=j+2}^{d} \mathcal{X}_i \right) \\ & = P_{V_{j+1}} \left(A_{(b_1, \ldots, b_j)} \right) = 0 \end{aligned}$$

and

$$\begin{aligned} & P_{V_1 \ldots V_j} \left(\left[A \times \prod_{i=j+2}^{d} \mathcal{X}_i \right]_{(b_{j+1}, \ldots, b_d)} \right) \\ & = P_{V_1 \ldots V_j} \left(A_{b_{j+1}} \right) = 0. \end{aligned}$$

Then

$$A_3 \times \prod_{i=j+2}^{d} \mathcal{X}_i \subseteq B.$$

Because P_V is non-singular, $P_V(B) = 0$, so

$$P_V \left(A_3 \times \prod_{i=j+2}^{d} \mathcal{X}_i \right) = P_{V_1 \ldots V_j V_{j+1}}(A_3) = 0.$$

Now, $A \subseteq A_1 \cup A_2 \cup A_3$ implies that

$$\begin{aligned} & P_{V_1 \ldots V_j V_{j+1}}(A) \\ & \leq P_{V_1 \ldots V_j V_{j+1}}(A_1 \cup A_2 \cup A_3) \\ & \leq P_{V_1 \ldots V_j V_{j+1}}(A_1) + P_{V_1 \ldots V_j V_{j+1}}(A_2) \\ & \quad + P_{V_1 \ldots V_j V_{j+1}}(A_3) \\ & = 0, \end{aligned}$$

so $P_{V_1 \ldots V_j V_{j+1}}(A) = 0$. Thus, by mathematical induction, for any positive integer, d, we have that $P_V \ll \mu$. □

Theorem 5.8.7. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a d-dimensional probability space and P_G a Bayesian factorization corresponding to DAG, G. If P is non-singular, then $P \ll P_G$ and the Radon-Nikodym derivative of P with respect to P_G, $\frac{dP}{dP_G}$, exists and $D(G) = \int_{\mathcal{X}} \log \frac{dP}{dP_G} dP$ is well-defined.*

Proof. To show that $P \ll P_G$, we use mathematical induction where each step is the number of factors in P_G. If P_G has one factor, then $P = P_G$, so $P \ll P_G$ in this case. Assume that $P \ll P_G$ for any factorization,P_G, having d factors. Let P_G^* have $d+1$ factors and $P_G^* = P_{G_1} P_{G_2}$ be some split of the factors with corresponding spaces $\mathcal{X}_{G_1} \times \mathcal{X}_{G_2} = \mathcal{X}$.

Assume $A \subseteq \mathcal{X}_{G_1} \times \mathcal{X}_{G_2}$ is measurable and $(P_{G_1} P_{G_2})(A) = 0$. Define $A_1 = \{x : P_{G_2}(A_x) > 0\} \times \mathcal{X}_{G_2}$, $A_2 = \mathcal{X}_{G_1} \times \{y : P_{G_1}(A_y) > 0\}$, and $A_3 = \{(x, y) : P_{G_1}(A_y) = P_{G_2}(A_x) = 0\}$. Notice that $A \subseteq A_1 \cup A_2 \cup A_3$.

From Fubini's theorem, we have that $0 = (P_{G_1} P_{G_2})(A) = \int_{\mathcal{X}_{G_1}} P_{G_2}(A_x) dP_{G_1}(x)$. Using Lemma 1.3.8 (Dembo 2019), $f \geq 0, \int f d\mu = 0 \Rightarrow \mu\{x : f(x) > 0\} = 0$, for the first equality, we must have

$$0 = P_{G_1}(\{x : P_{G_2}(A_x) > 0\}) = P(\{x : P_{G_2}(A_x) > 0\} \times \mathcal{X}_{G_2}) = P(A_1). \quad (5.162)$$

Using the same construction but switching G_1 and G_2, we also have that $0 = P(A_2)$. $P(A_3) = 0$ follows from the definition of non-singular. This shows that $P \ll P_{G_1} P_{G_2} = P_G^*$. Thus, $P \ll P_G$ for any number of factors.

Now, we may apply the RN theorem, so there exists a measurable function, f such that for any measurable set $A \subseteq \mathcal{X}$, $\int_A f dP_G = P(A)$ and f is unique almost everywhere $[P]$. Using lemma 7.4 (Gray Entropy), $D(G) = \int_{\mathcal{X}} \log \frac{dP}{dP_G} dP$ is well-defined. □

Theorem 5.8.8. *Let $(\mathcal{X}, \mathcal{B}, P)$ be a non-singular d-dimensional probability space*

such that for each $i \in [d]$, there exists a $\mathcal{B}_i$-measurable set $E \subseteq \mathcal{X}_i$ such that $P_i \ll \lambda$ on E and $\mathcal{X} \backslash E$ is nowhere dense $[\lambda]$ in $\mathcal{X}$. Let P_G be a Bayesian factorization with respect to DAG, G. If $\frac{dP}{dP_G}$ is the Radon-Nikodym derivative of P with respect to P_G, then

$$\frac{dP^{(x)}(r)}{dP_G^{(x)}(r)} \to \frac{dP}{dP_G}(x) \tag{5.163}$$

almost surely $[P]$ on the support of P.

Proof. Recall that for $W \in [d]$, $x \in \mathcal{X}$, and $r > 0$,

$$B_W^{(x)}(r) := \{(y_i : i \in [d]) \in \mathcal{X} : |x_i - y_i| \leq r, i \in W\} \tag{5.164}$$

$$P_W^{(x)}(r) = P\left(\{(y_i : i \in [d]) : |x_i - y_i| \leq r, i \in W\}\right). \tag{5.165}$$

Using the dominating measure (mix of counting and Lebesgue) from lemma 5.5.1, we first show that $\frac{P^{(x)}(r)}{\mu^{(x)}(r)} \to \frac{dP}{d\mu}$.

Let $x \in \mathcal{X}$, $C^{(x)} = \left\{i \in [d] : P_i^{(x)}(0) = 0\right\}$ and $D^{(x)} = \left\{i \in [d] : P_i^{(x)}(0) > 0\right\}$. Because the set of discrete points is nowhere dense, there must be a $\zeta > 0$ such that for each $y \in B_{C^{(x)}}^{(x)}(\zeta)$, $P_{C^{(x)}}^{(y)}(0) = 0$. Assume $C^{(x)} \neq \varnothing$. Let λ be the $\left|C^{(x)}\right|$-dimensional Lebesgue measure. Then $P \ll \lambda$ on $B_{C^{(x)}}^{(x)}(\zeta)$. Using Rudin thm 7.8 (Lebesgue theorem),

$$\frac{P_{C^{(x)}}^{(x)}(r)}{\lambda^{(x)}(r)} \to \frac{dP_{C^{(x)}}}{d\lambda}(x) \tag{5.166}$$

as $r \to 0$ almost everywhere $[\lambda]$. Note: $C^{(x)}$ or $D^{(x)}$ may be empty; define $P_{\varnothing}(A) = 1$. Further, if x has more than $|C^{(x)}|$ coordinates; the discrete coordinates of x indicate location for the continuous density.

Claim:

$$f(x) := P_{D^{(x)}}^{(x)}(0) \frac{dP_{C^{(x)}}}{d\lambda}(x) = \frac{dP}{d\mu}(x). \tag{5.167}$$

Partition $\mathcal{X}$ into its uniformly continuous regions according to P: x and y are in the

same partition if $C^{(x)} = C^{(y)}$ and there exists a subset $B \subseteq \mathcal{X}$ with $x, y \in B$, such that $\frac{dP_{C^{(x)}}}{d\lambda}(x)$ is uniformly continuous on B. Let $\mathcal{T}$ be this partition of $\mathcal{X}$ described above. Points $x \in \mathcal{X}$ such that $D^{(x)} = [d]$ (all coordinates are discrete), will be singleton classes $\{x\} \in \mathcal{T}$. It is clear that this defines an equivalence class for points in $\mathcal{X}$. Further, there can be, at most, countably many classes in this partition.

Let $A \subseteq \mathcal{X}$ and μ be the dominating measure (mix of counting measure and Lebesgue). Below, λ is the $|C^{(x)}|$-dimensional Lebesgue measure, which does not change within each partition; we suppress the dimension. Then

$$\begin{aligned}
\int_A f(x)d\mu(x) &= \sum_{T\in\mathcal{T}} \int_{A\cap T} f(x)d\mu(x) && (5.168)\\
&= \sum_{T\in\mathcal{T}} \int_{A\cap T} P^{(x)}_{D^{(x)}}(0)\frac{dP_{C^{(x)}}}{d\lambda}(x)d\mu(x) && (5.169)\\
&= \sum_{T\in\mathcal{T}} I_{A\cap T}(x)P^{(x)}_{D^{(x)}}(0)\int_{A\cap T}\frac{dP_{C^{(x)}}}{d\lambda}(x)d\lambda(x) && (5.170)\\
&= \sum_{T\in\mathcal{T}} I_{A\cap T}(x)P^{(x)}_{D^{(x)}}(0)P_{C^{(x)}}(A\cap T) && (5.171)\\
&= \sum_{T\in\mathcal{T}} P(A\cap T) = P(A) && (5.172)
\end{aligned}$$

Because P_G is the product of regular conditional probabilities of P, the set of discrete points for each dimension must be the same for both. Using the same proof for P, we have

$$P^{(x)}_{G,D^{(x)}}(r)\frac{P^{(x)}_{G,C^{(x)}}(r)}{\lambda^{(x)}(r)} \to P^{(x)}_{G,D^{(x)}}(0)\frac{dP_{G,C^{(x)}}}{d\lambda}(x) = \frac{dP_G}{d\mu}(x) \tag{5.173}$$

as $r \to 0$.

Because P is non-singular, $P \ll P_G$, so $\frac{dP}{dP_G}$ exists. Then, for support points of P, using properties of RN derivatives

$$\frac{P^{(x)}(r)}{P^{(x)}_G(r)} = \frac{P^{(x)}_{D^{(x)}}(r)P^{(x)}_{C^{(x)}}(r)}{P^{(x)}_{G,D^{(x)}}(r)P^{(x)}_{G,C^{(x)}}(r)} \to \frac{dP}{dP_G}(x) \tag{5.174}$$

as $r \to 0$ almost surely $[P]$. $\square$

Chapter 6

American Men's Internet Survey

6.1 Introduction

With the advent and increasing use of HIV Pre-exposure prophylaxis (PrEP), drugs that effectively reduce risk of HIV acquisition, the established understanding for HIV and STI prevention may be shifting. Sadly, HIV and other sexually transmitted infections (STI) incidence disproportionately affect young men who have sex with men (MSM) of color who are less likely know about or able to afford PrEP [85]. This phenomenon, along with many others, can obfuscate the larger picture of the relationships between interacting risk factors leading to HIV and STI infections.

A clearer picture could aid the development of new interventions for HIV and STI reduction, particularly in hard-hit communities. In this study, we present a new data method for depicting causal relationships between variables developed specifically for the kinds of data found in epidemiological work. With it, we explore causal relationships found on select variables within a randomized subset of responses from the American Men's Internet Survey (AMIS). We examine these relationships with generalized linear regression models on the remaining data to gain a clearer idea of risk factors leading to STI acquisition among men who have sex with men.

6.2 Background

PrEP is a class of HIV antiretroviral drug combinations intended to prevent an HIV infection in HIV-negative individuals considered to be high-risk, such as men who have sex with men (MSM), trans women, serodiscordant, sexually-active couples, and persons who inject drugs. In 2012, the US Food and Drug Administration (FDA) approved Truvada (tenofovir disoproxil fumarate and emtricitabine) [86]as the first drug combination for use as PrEP. Truvada was originally develop to treat HIV and is still used in combination with another antiretroviral drug to treat HIV. In 2019, the FDA also approved Descovy (tenofovir alafenamide and emtricitabine) for use as PrEP [87].

PrEP is very effective for preventing HIV infection in at-risk individuals who follow adherence guidelines [88]. With the burgeoning use of PrEP, particularly among men who have sex with men (MSM), some researchers have observed risk compensation, associated with the reduction in HIV-acquisition risk, leading to an increase in condomless sex [89, 90, 91]. While PrEP decreases the risk of HIV infection, it does not protect against other STIs such as syphilis, gonorrhea, chlamydia, or HPV. Consequently, PrEP's potential to shift sexual risk behaviors may be disrupting established causal pathways leading to HIV and STIs. Moreover, the racial disparity in PrEP uptake in the US, along with the historic disparity in HIV incidence [92], further complicates the causal landscape.

Causal discovery, the estimation of causal pathways through the variables of a dataset with a diagram (called a Bayesian network or directed acyclic graphs (DAG)), is rare in clinical and epidemiological literature. This may be due, in part, to the difficulty existing methods have with mixed data. However, the benefit of such models in this case is clear: a causal diagram could illuminate the role that PrEP and race, for example, play in potentially confounding multiple behaviors and outcomes. In this study, we use the 2019 American Men's Internet Survey (AMIS) [93], of demographic information, risk behaviors and outcomes for 10,000 MSM in the US. The study team at Emory University responsible for AMIS shared these data with IRB approval (IRB ID: STUDY2019_00000572).

This chapter aims to present preliminary results; however, our larger aim is two fold: (1) graphically estimate the underlying causal pathways through the data and (2) make causal discovery accessible for epidemiologists. Casual discovery provides a data-driven approach to organizing data and informing variable selection. Regression provides a fine-tuned look at controlled (adjusted) associations with an outcome. Together, these approaches provide a more complete understanding of the causal relationships underlying data. Connecting causal Bayesian network models and regression, we will take half the observations to estimate the graphical model,

then use it to inform regression model selection for outcomes of interest on the other half. Ideally, this work will provide evidence for potential public health interventions and further research while giving visibility to and appreciation for a new data method.

6.3 Methods

The AMIS data is collected annually through an internet survey targeting MSM through website ad banners and emails through website membership. Part participants who agreed to be contacted were also invited to participate again. To take the survey, participants must be (1) older than 15 years, (2) cisgender male, (3) living in the US, and (4) report having either oral or anal sexual contact with a male partner in the past or identify as gay or bisexual.

Because the goal of this work is to explore the relationship between PrEP use, demographic information, risk behavior and outcomes, we selected participant responses based on the following: (1) responded being HIV-negative, (2) reported oral or anal sexual contact with at least one male partner in the past 12 months, and (3) had no missing values among the variables we chose. We included only HIV-negative individuals because PrEP is not recommended for individuals living with HIV. We required one or more oral or anal sex partners within the past 12 months to screen those not at risk for HIV/STI infection. Last, we included those with no missing or unknown responses among the variables used as is customary with regression. We included the following variables:

- `AGE`: Age in years at time of survey
- `RACE`: Race as Hispanic, Black, Non-Hispanic White, and Other (we collapsed categories with few responses)
- `EDU`: Level of education completed (EDU) as an integer 1 through 4

- `INCOME`: Household income, as an integer 1 through 4
- `USED_PREP`: PrEP use in the past 12 months
- `NUM_PARTNERS`: Number of male, oral or anal sexual partners as in integer
- `CONDOMLESS`: Number of condomless, anal or oral sex, male partners

We randomly assigned individual responses either to group 1 or 2; we used group 1 data for estimating the Bayesian network, and group 2 for the following regression. We put each group's data into separate files to prevent accidental usage of the same individual response for the different methods. We first estimated the Bayesian network on group 1 data to develop hypotheses about causal pathways among the variables. We then chose regression models to explore the findings from the Bayesian networks. We used generalized linear models for all regression models because of their familiarity and widespread use in epidemiological work. However, it should be noted that the Bayesian network estimation method does not assume linear, statistical associations between variables, as generalized linear models do. Therefore, discrepancies between the Bayesian network and regression models may be due, at least in part, to this difference.

For causal discovery, we used mixed graph divergence from Chapter 5 combined with a variation of greedy equivalence search (GES) [4] and greedy interventional equivalence search (GIES) [82]. A *Bayesian network* is a graph structure (a diagram) where variables are nodes and the arrows connecting the nodes are called edges or directed edges, indicating a causal association. Each edge in the Bayesian network is directed and following the edges cannot lead to a cycle, which is why Bayesian networks are also called directed acyclic graphs (DAG). Graph divergence quantifies the information-theoretic distance between any given Bayesian network on the variables of a distribution and its true distribution. In theory, a graph divergence of zero indicates that a Bayesian network perfectly fits the data. Due to estimation error from the variability in the data, it is very unlikely that the true

Bayesian network will have an estimated graph divergence (EDG) of exactly zero, but it should be close, as with all estimation. However, it is possible to logically determine particular edges which, if added to the true causal network, will retain a zero graph divergence on the augmented network. In contrast, removing one or more edges from the true causal network will always result in a greater graph divergence. Because causal networks are typically thought to be sparse, the number of possible networks with a given number of edges increases with that number of edges. This potentially increases the risk of false-positive edges in near-zero graph divergence networks. For these reasons, we prefer causal networks with a small, but positive graph divergence estimate and fewer edges. Determining a balance between edge count and near-zero divergence is left to future work.

While graph divergence quantifies the fit of a causal network to a dataset, it does not provide a way to search through the possible causal networks. To do this, we implemented versions of GES [4] and GIES [82] tailored to the computationally intensive needs of estimating graph divergence. The algorithm uses three general types of step, forward, backward, and tuning, to traverse though the space of Bayesian networks on the variables provided. The tuning step is only in GIES. Each step inherits the Bayesian network chosen by the last step. The algorithm is greedy, a computer science term, because at each step, it chooses to pass on the Bayesian network that maximizes the EGD among nearby Bayesian networks. A *forward* step estimates the graph divergence on all possible Bayesian networks resulting from *adding one directed edge* to the inherited network. A *backward* step estimates the graph divergence on all possible Bayesian networks resulting from *deleting one edge* from the inherited Bayesian network. A *tuning* step estimates the graph divergence on all possible Bayesian networks resulting from *switching direction of one edge* on the inherited graph. The algorithm takes forward steps until it can no longer improve the EGD from the previous step. Analogously, the algorithm takes backward/tuning steps until it can no longer improve the EGD from the previous step. When taking

the same type of step no longer improves the EGD, it changes to backward from forward, tuning from backward, and forward from tuning. The algorithm begins with a forward step on the network with no edges. It ends when none of the steps can improve the EGD score. We engineered all code in Python 3.7.7. (The code is not currently ready for public use, but the package will be made publicly available.)

GES and GIES typically use a composite Bayesian information criterion (BIC) score to estimated graph fit, which penalizes Bayesian networks with many edges. Graph divergence does not. However, unlike BIC, a graph divergence of zero is at least theoretically significant as previously discussed. Thus, there may be some benefit to using graph divergence over BIC, though more research is required. To avoid spending time on over-fitted networks (with too many edges), we used the absolute value of EDG for the algorithm. Graph divergence is non-negative in theory but the estimator can give negative values. In this work, we presents the Bayesian network with optimal absolute EGD between the two networks whose scores flank zero once the algorithm finds a network with a negative score. In the future, we would like to assess this with a statistical test. As of this draft, we did not aggregate Markov equivalent Bayesian networks. We used Graphviz to display networks.

6.4 Preliminary Results

The full dataset included 10130 total responses. Of those, 3469 indicated either living with HIV, did not know, or preferred not to respond. 6661 responses indicated being HIV-negative. Of those, 895 reported having zero oral or anal, male sex partners; 5766 reported one or more. Among those with at least one oral or anal sex partner in the past 12 months, 1216 did not know or preferred not to respond to at least one of the of questions used in the analysis. The screening left a total of 4550 individual responses. The number of included/excluded individual responses are shown in figure 6.1.

Table 6.1 gives basic counts and means for each variable used in this analysis

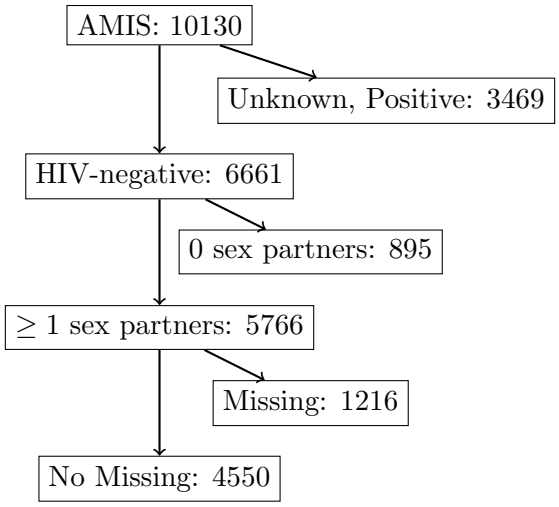


Figure 6.1: The flowchart above shows the number of individual responses that were included and excluded at each step.

among the screened 4550 responses. The median respondent was 29 years at the time of the survey. The respondents were majority white (63%), and mostly possessed a college degree or greater (57%) with varied household incomes. Twenty-seven percent used PrEP within the past 12 months. Within the 12 month prior to taking the survey, the median participants reported having one condomless, oral or anal, male sex partners, out of five total oral or anal, male sex partners. These medians and proportions remained steady after randomization into groups 1 and 2 with 2227 responses randomly selected for group 1 and 2323 for group 2.

6.4.1 Causal Discovery

As discussed in Section 6.3, we used a heuristic to select estimated Bayesian networks. Figure 6.2 shows EGD scores for all networks the algorithm traversed. Initially, the algorithm took 12 steps forward; it evaluated some 13-edge networks, but none improved absolute EDG. Stepping backward yielded no networks that improved EDG. As this was somewhat expected, we altered the algorithm to allow backward steps if the optimal graph (removing one edge) increased EDG by a spec-

	Total	Group 1	Group 2
n	4550	2227	2323
Age (years)	29 (24,42)	29 (24,42)	29 (24,42)
RACE			
Black	768 (17%)	377 (17%)	391 (17%)
Hispanic	588 (13%)	280 (13%)	308 (13%)
Non-Hispanic White	2872 (63%)	1408 (63%)	1464 (63%)
Other	322 (7%)	162 (7%)	160 (7%)
EDU			
Less High School	52 (1%)	22 (1%)	30 (1%)
Some High School	15 (0%)	11 (0%)	4 (0%)
High School	223 (5%)	110 (5%)	113 (5%)
Associates	1678 (37%)	848 (38%)	830 (36%)
College or more	2582 (57%)	1236 (56%)	1346 (58%)
Yearly Household Income ($1000)			
0 to 19	566 (12%)	275 (12%)	291 (13%)
20 to 39	926 (20%)	442 (20%)	484 (21%)
40 to 74	1408 (31%)	704 (32%)	704 (30%)
75 or more	1650 (36%)	806 (36%)	844 (36%)
Used PrEP (past 12 mos)	1111 (24%)	528 (24%)	583 (25%)
STI Test (past 12 mos)	2704 (59%)	1312 (59%)	1392 (60%)
Total Sex Partners (12 mos)	5 (2,10)	5 (2,10)	5 (2,10)
Condomless Sex Partners (12 mos)	1 (1,3.8)	1 (1,3)	2 (1,4)

Table 6.1: The table above shows the median (interquartile range) for variables with more than 4 unique values, and counts (percentage) otherwise, stratified by randomized group.

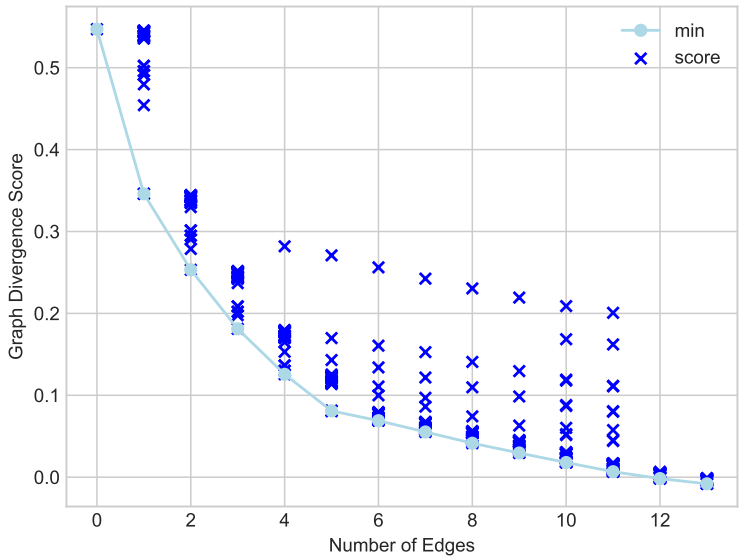


Figure 6.2: Above is a scatter plot of estimated graph divergence by number of edges for all Bayesian networks within the algorithm search path.

ified allowance. We decreased the allowance as the algorithm progressed from 0.05 to to 0.001 over ten backward iterations. We did this to encourage the algorithm to consider a more diverse set of graphs.

After several iterations, it became clear that age and race were very statistically dependent in the data; these were the first variables connected at the beginning of the algorithm and never disconnected throughout the algorithm's progression. This association is probably induced via sampling bias. Because it seems unlikely or impossible for age and race to influence each other, or for other variables to influence age and race, we decided to only consider Bayesian networks where age and race had no incoming arrows. To enforce this in the algorithm, we assigned a score value of infinity to any Bayesian network with incoming arrows to race and age.

Figure 6.3 shows the optimal Bayesian network according to the criteria we set, with an estimated graph divergence of -0.001685. The optimal 11-edge network scored 0.006733; it did not include an edge between `HOUSEHOLD_INCOME` and `USED_PREP`.

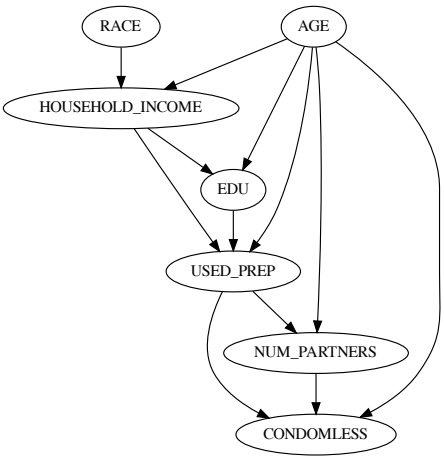


Figure 6.3: Optimal 12-edge Bayesian network

Some of the edges in figure 6.3 would not be oriented using Markov equivalence. For example, the edge between `HOUSEHOLD_INCOME` and `EDU` could probably be oriented in either direction. We kept the orientation given by the algorithm for transparency. In general, Bayesian networks are said to be Markov equivalent if they contain the same conditional independence relationships between variable, and are causally indistinguishable causally without stronger assumptions. Reference [6, Chapter 3] provides a clear treatment of Markov equivalence and Bayesian networks.

6.4.2 Regression

We ran three regression models partly based on the estimated Bayesian networks from Section 6.4.1. We chose `USED_PREP`, `NUM_PARTNERS`, and `CONDOMLESS` as outcome variables for the regression models. We used logistic regression for the `USED_PREP` model and linear regression for the `NUM_PARTNERS` and `CONDOMLESS` models in R Programming Language. We included `AGE`, `RACE`, `EDU`, and `HOUSEHOLD_INCOME` in each models. We included outcome variables in the regression models according to their location within the Bayesian networks when it made clinical sense. More specifi-

	USED_PREP		NUM_PARTNERS		CONDOMLESS	
	log odds	p-value	mean change	p-value	mean change	p-value
AGE per 10 yrs	0.08	0.033	0.61	0.065	0.07	0.518
RACE: Black	0.44	0.001	-0.11	0.921	-0.01	0.986
RACE: Hispanic	-0.11	0.516	4.03	0.002	0.19	0.672
RACE: Other	0.39	0.029	3.88	0.016	-1.13	0.041
EDU	0.47	0.000	0.11	0.857	-0.10	0.621
INCOME	0.11	0.037	-0.22	0.603	0.17	0.247
USED_PREP	—	—	9.63	0.000	2.52	0.000
NUM_PARTNERS	—	—	—	—	0.31	0.000

Table 6.2: The table above gives generalized linear regression parameter estimates based on the causal discovery model.

cally, CONDOMLESS (the number of condomless partners), would seems to depend on NUM_PARTNERS. We chose outcomes as independent variables for these models with this in mind. The causal ordering for USED_PREP with respect to the other outcomes is less clear; however, the Bayesian network in figure 6.3 indicates a proximity to independent variables, which is why we included it as in independent variable in the other models. Table 6.2 give the regression results for each model.

Comparing the regression models to the estimated Bayesian networks, we see some agreement and some discrepancies. Perfect agreement would indicate that, for example, within the USED_PREP model, AGE, HOUSEHOLD_INCOME, and EDU would be significant, while all other variables would not be. For the CONDOMLESS model to agree with the estimated Bayesian networks, we would expect only NUM_PARTNERS, AGE, and USED_PREP to be significant, which is not the case. In general, perfect agreement in this setting means that only neighboring variables in the Bayesian network will be significant in the regression, assuming we do not include a downstream cause of the outcome in the regression. In much of the epidemiology literature, using a significance level of $\alpha = 0.05$ is fairly standard. Instead, we will use a significance level of $\alpha = 0.01$ which is more conservative but not as conservative as a Bonferroni correction for 21 tests.

There are several possible reasons for discrepancies between the regression mod-

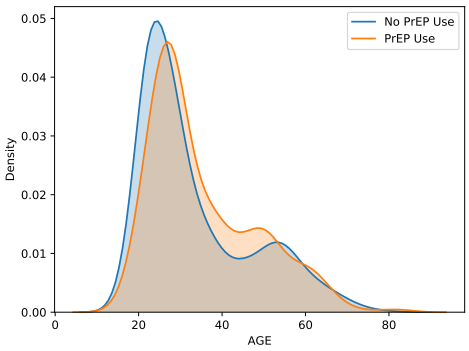
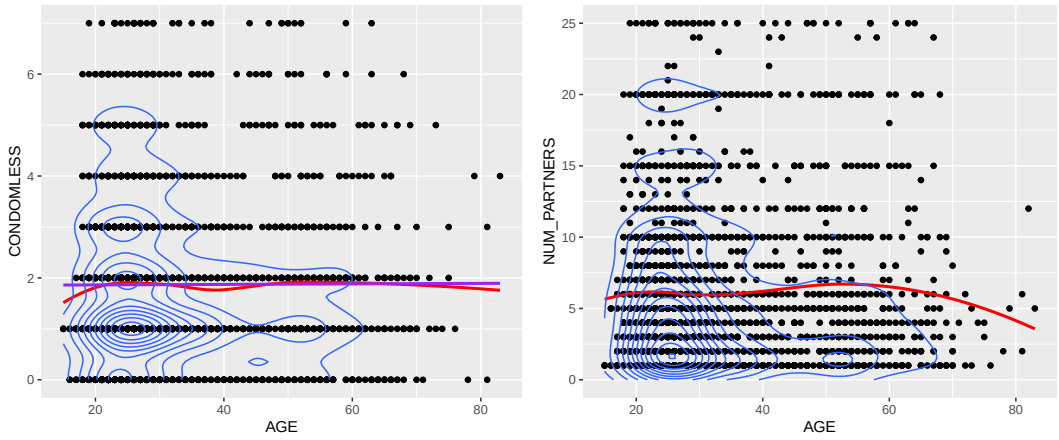


Figure 6.4: Density Estimates for `AGE` by `USED_PREP`. Wilcoxon rank sums test p-value $= 4.6 \times 10^{-8}$, t-test p-value $= 0.004$

els and Bayesian network. From a data science point of view, we would not expect perfect agreement because of differing model assumption. Generalized linear models test for *mean* differences between a variable (or indicator, as in `RACE`: Black) and the baseline category assuming all relationships are linear. Graph divergence, the statistic used to estimate Bayesian model fit, is non-parametric in that it detects more than mean difference and does not assume linear relationships. Consider the relationship between `USED_PREP` and `AGE`: the Bayesian models indicate that the other variables in the data cannot explain their relationship, while the `PREP_USED` regression model indicates if a relationship exists, it can be explained by the other variables. Density estimates of `AGE` by `USED_PREP` clearly illustrate this; figure 6.4 shows subtle differences between these distributions. Moreover, the non-parametric, Wilcoxon rank sums test gives a p-value of 4.6×10^{-8} while a t-test gives a p-value of 0.004, perhaps indicating information is being lost with parametric models.

Similarly, the relationships between `NUM_PARTNERS` and `CONDOMLESS` both with `AGE` may be missed by linear models as well. Figure 6.5 show these relationships with 2D density plots: the more central contour curves indicate greater density. In both of these plots, the mean of both `NUM_PARTNERS` and `CONDOMLESS` remain fairly constant over varying values of `AGE`, but its clear that in both, the variance of



(a) CONDOMLESS vs AGE (b) NUM_PARTNERS vs AGE

Figure 6.5: 2D Density Plots (blue contours) with local regression (red curve). There were responses for both plots on the y-axis great than 25. The plots only show responses below 25 to better illustrate the densities and mean curves

NUM_PARTNERS and CONDOMLESS tend to decrease with AGE. This type of association would not likely be captured with linear models but should with non-parametric models.

Though, typically, we would expect the opposite to happen more frequently: significant, controlled associations between independent variables and outcomes in regression models not appearing to be connected in the Bayesian network. As briefly explained in Section 6.3, networks with many edges (relative sample size and variables) have a greater chance of containing false positive edges. False positive edges can cause directional errors with surrounding edges. For these reasons, it is typically preferred to present Bayesian networks that include less edges, with false negative associations preferred over false positives. Fortunately, we would expected association strength to affect the appearance of an edge in a Bayesian network, with stronger associations more likely to be present. Interestingly, much scientific research, knowingly or otherwise, gravitates in the opposite direction, toward preferring false negative associations over false negative due to the widespread failure to control for multiple testing.

Another possible reason for the absence of an edge in a non-parametric, estimated Bayesian network compared to a generalized linear model can again be due to the fact that graph divergence does not assume all relationships are linear. Though we see no evidence of this phenomenon occurring here, it is possible. For example, assume there are three variables, A, B, and C with $A \to B \to C$ where the arrow indicates causal association. If $A \to B$ and $B \to C$ have non-linear relationships but a linear relationship does exist between A, and C, a generalized linear model may not be able properly control for B in this setting.

The last reason we give for this type of discrepancy is the difference in treatment of categorical variables between regression and the graph divergence statistic within the Bayesian network estimation. Regression models handle categorical variables by establishing a baseline subgroup (Private for `RACE`, for example) and creating indicator variables for all other groups. While this can be done for the Bayesian network as well, the additional variables in the could lead to more error. Though this versatility might be helpful in some settings including this one. Graph divergence calculates an estimate for each observation point then averages over all observation estimates. Each category is represented in the overall estimate proportional to its makeup in the data. Because those non-Whites make up less than 40% of the data, any differences within this group will represent less than 40% of the graph divergence estimate. This may be why `RACE: Black` is significant in the `USED_PREP` model but not reflected in the Bayesian network.

6.5 Discussion

The estimated Bayesian network in figure 6.3 communicates that, age, more than any other risk factor, influences much of the causal landscape, including the input variables level of education, and income. This seems very reasonable, and something not evaluated by the regression models. Age also seemed to influence all outcomes, PrEP use, number of total and condomless oral/anal sex partner directly. Figure 6.4

shows a complex marginal relationship between PrEP use and age, that may have been missed by the PrEP regression model, with more younger (15 to early 20s) and older (70 and up) non-PrEP users but more PrEP users between those age groups. Similarly, there is more variation in number of sex partners with and without a condom among younger participants, even though the mean number of partners does not depend as much on age. As such, there may be a greater need for STI testing among younger MSM.

Race is only connected with income in the Bayesian network. The regression models indicate that compared to Non-Hispanic Whites, Blacks are more likely to use PrEP. It also shows that Hispanics, compared to Non-Hispanic Whites, have on average four more sex partners on average, though the number of condomless sex partners is similar to Non-Hispanic Whites. It is worth noting that health disparities for HIV and STI acquisition and treatment do exist, typically with the Black and Latino communities disproportionately negatively affected [94, 95]. Race is probably not connected with other variables in the estimated network due to the fact that the population taking this survey is majority Non-Hispanic White with all other groups making up less than 40% of the survey population and represented proportional in graph divergence estimator and not directly compared with Non-Hispanic Whites.

Level of education appears to influence PrEP use. This is clear from all models presented in this work. We read the connection between PrEP use and level of education as probably causal (or possibly confounded) with education affecting PrEP use. The regression model indicates that those reporting higher levels of education were more likely to report PrEP use as well. Additionally, the Bayesian network shows that education is affected by both age and household income, again, not seen with the regression models.

The Bayesian network and the regression models both indicate a connection between PrEP use and number of total and condomless partners. Moreover, both models agree that the relationship between PrEP use and number of condomless

partners reported is not mediated by total number of partners. Because physicians would probably recommend PrEP to a patient reporting having condomless sex with multiple partners within a short period of time, there is likely a causal feedback loop from condomless partners to PrEP uses. However, there is likely forward causation from PrEP use to condomless sex as well. Though this conclusion does not follow exclusively from the Bayesian network nor regression models though they both give a consistent narrative. Using PrEP does indicate that a patient and physician think there is at least some risk of condomless sex and HIV acquisition in the future for the patient. It is plausible that the reduction in HIV-acquisition risk may actually result in reduced pressure to use condoms, after using PrEP. However, this assumes that there is not latent confound influencing both PrEP use and number of sex partners. While this can be debated, other studies have also found a drop in condom use corresponding with the rollout of PrEP [96, 97].

Regardless of causal direction, it is clear that PrEP users compared to those not on PrEP have more condomless sex partners. The CDC already recommends that PrEP users be tested every three months; a recommendation supported by these finds. Further, the incidence of STIs within the PrEP using population may be representative of a wider, less-screened, at-risk population. Local centralized reporting on all test outcomes, including negative tests, among PrEP users may help public health officials accurately determine the occurrence of an outbreak more quickly. The cost of potentially lost privacy and data breaches would need to be taken into account before such a program were set up, but such analysis is beyond the scope of this work.

Causal discovery together with regression modeling can give a more complete picture of interacting risk factors leading to outcomes. In general, causal discovery could be a useful exploratory data analysis tool to help epidemiologists solidify potential causal pathways prior to more formal analysis with regression modeling. This type of approach may aid researchers locate new and interesting research ideas

that can be further fine-tuned with regression models. This work remains at an early stage, but we hope it demonstrates a more general approach to statistical modeling that benefits both from the data-driven causal model and expert understanding, to improve scientific research.

Chapter 7

Conclusion and Future Work

The aim of this thesis was to provide a usable data analysis tool to help researchers, with epidemiologists in mind, explore causal structure within datasets for mixed data without making parametric assumptions. While this work is ongoing, I believe I have made progress toward that goal.

Chapter 3 was my Ph.D. qualifier and my first attempt at applying a causal discovery and writing software. This experience was very informative for learning how researchers tend to interact with graphical models in practice. In this setting, graphs with fewer, carefully chosen nodes tend to be more helpful. Smaller graphs are easier to understand. Unrelatedly, many researchers are accustomed to parameter estimates, p-values, and confidence intervals. Given this, it may be more helpful to combine graphical models and regression. This paper was published in PLoS One [98].

Chapter 4 develops a method for estimating conditional mutual information on mixed data. This paper was my first attempt writing a theory-oriented paper. In it, we showed that our estimator was theoretically consistent and that it performed better than other, estimators. I used this paper as an opportunity to learn Python and make a publicly, usable software package. We submitted this paper to IEEE Transactions on Information Theory in December 2019. We were asked to revise

and resubmit.

Originally, we planned to use CMI combined with the PC algorithm to do structure learning. But, because we did not know the distribution of the CMI estimator, it would have been necessary to devise an algorithm that did not require p-values. It was this that led problem that led to the following chapter.

Chapter 5 is ongoing work. Rather than using many CMI calculations to estimate a graphical structure using the PC algorithm, this paper exploits that fact that for any DAG, it is possible to decompose its conditional independence relationships with entropy. We developed an estimator that could also handle mixed data using what we learned from Chapter 4.

For this paper, Cosma, Larry, and I set out to prove that graph divergence estimator obeyed the central limit theorem (CLT). At the start, I did not understand how challenging this would turn out to be. My first mistake was attempting to prove the CLT without smoothness assumptions and with the mixed and continuous distributions combined. Proving that this estimator is consistent does follow directly from the consistency proofs for CMI. In a higher-dimensional setting, it seemed important to understand how dimension affects mean squared error (MSE). This was not included in Chapter 4.

To make the problem easier, I focused only on continuous, Hölder-smooth densities, planning on adding in the discrete part later. The discrete problem was largely already solved in reference [78]. I am currently working on proving the MSE convergence rate. We will try to publish this work as a standalone paper and publish the mixed paper separately.

Chapter 6 is meant as a preliminary look at the AMIS data, written for this thesis but we will not try to publish it as written here. Before we publish a paper on the AMIS data, I would like to better understand how to use the graph divergence estimator with a GES-like algorithm in light of the fact that it does not penalizes less sparse graphs. Fortunately, the graph we presented in Chapter 6 does seem to make

sense, and generally agrees with the regression models. But, having this problem figured out will make the paper much stronger. I would also like to work more closely with the group of researchers at Emory University responsible for collecting the AMIS data. Their knowledge of the field as well as the data will surely improve the paper quality.

For the future, assuming we will be able prove that the graph divergence estimator obeys the CLT, I think it could be very useful to explore inference on Bayesian networks. In particular, a graph divergence CLT could allow for confidence sets of graphs. With it, we could provide edge probabilities to help researchers gauge the likelihood of a causal association from the graphs themselves.

Alternatively, methods for estimating causal graphical models on longitudinal, or time-series data would also be very interesting. This could be of value in fields with feedback loops. One possible avenue could begin by expanding on the CMI estimator in [99] to estimate conditional information rate [100, 101] and adapting a causal graph estimation method appropriately.

Another natural direction of interest is causal discovery on data with missing values. Because the estimation process is already iterative, an EM-algorithm-based approach may be helpful, using the current structure to impute missing values then re-estimate the structure on the imputed data.

Finally, my research history points to the synergy that can occur when application informs theory and visa versa. Using this paradigm for my dissertation research likely contributed to me being awarded the Ford Foundation Dissertation Fellowship. Though nearly all of my applied research has been in infectious disease, I am open to expand moving forward. But, regardless of the direction, it is important to me that my work contributes to overall social progress.

Bibliography

[1] R. W. Robinson, "Counting unlabeled acyclic digraphs," in *Combinatorial mathematics V*. Springer, 1977, pp. 28–43.

[2] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson, *Causation, prediction, and search*. MIT press, 2000.

[3] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the pc-algorithm," *Journal of Machine Learning Research*, vol. 8, no. Mar, pp. 613–636, 2007.

[4] D. M. Chickering and C. Meek, "Finding optimal bayesian networks," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 94–102.

[5] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour, "Generalized score functions for causal discovery," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1551–1560.

[6] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[7] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, no. Oct, pp. 2003–2030, 2006.

[8] T.-W. Lee, "Independent component analysis," in *Independent component analysis*. Springer, 1998, pp. 27–66.

[9] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[11] S. L. Lauritzen, "Graphical models, volume 17 of oxford statistical science series," 1996.

[12] O. Banerjee, L. E. Ghaoui, and A. dAspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *Journal of Machine learning research*, vol. 9, no. Mar, pp. 485–516, 2008.

[13] N. Meinshausen, P. Bühlmann *et al.*, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[14] B. Fellinghauer, P. Bühlmann, M. Ryffel, M. Von Rhein, and J. D. Reinhardt, "Stable graphical model estimation with random forests for discrete, continuous, and mixed variables," *Computational Statistics & Data Analysis*, vol. 64, pp. 132–152, 2013.

[15] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, pp. 417–473, 2010.

[16] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[17] P. D. Wadhwa, J. F. Culhane, V. Rauh, S. S. Barve, V. Hogan, C. A. Sandman, C. J. Hobel, A. Chicz-DeMet, C. Dunkel-Schetter, T. J. Garite *et al.*, "Stress, infection and preterm birth: a biobehavioural perspective," *Paediatric and Perinatal Epidemiology*, vol. 15, no. s2, pp. 17–29, 2001.

[18] N. Dole, D. A. Savitz, I. Hertz-Picciotto, A. M. Siega-Riz, M. J. McMahon, and P. Buekens, "Maternal stress and preterm birth," *American journal of epidemiology*, vol. 157, no. 1, pp. 14–24, 2003.

[19] R. L. Copper, R. L. Goldenberg, A. Das, N. Elder, M. Swain, G. Norman, R. Ramsey, P. Cotroneo, B. A. Collins, F. Johnson *et al.*, "The preterm prediction study: Maternal stress is associated with spontaneous preterm birth at less than thirty-five weeks' gestation," *American Journal of Obstetrics and Gynecology*, vol. 175, no. 5, pp. 1286–1292, 1996.

[20] M. S. Kramer, J. Lydon, L. Séguin, L. Goulet, S. R. Kahn, H. McNamara, J. Genest, C. Dassa, M. F. Chen, S. Sharma *et al.*, "Stress pathways to spontaneous preterm birth: the role of stressors, psychological distress, and stress hormones," *American journal of epidemiology*, vol. 169, no. 11, pp. 1319–1326, 2009.

[21] R. L. Goldenberg, J. F. Culhane, J. D. Iams, and R. Romero, "Epidemiology and causes of preterm birth," *The lancet*, vol. 371, no. 9606, pp. 75–84, 2008.

[22] L. M. Glynn, C. D. Schetter, C. J. Hobel, and C. A. Sandman, "Pattern of perceived stress and anxiety in pregnancy predicts preterm birth." *Health Psychology*, vol. 27, no. 1, p. 43, 2008.

[23] G. E. Miller, J. Culhane, W. Grobman, H. Simhan, D. E. Williamson, E. K. Adam, C. Buss, S. Entringer, K.-Y. Kim, J. F. Garcia-Espana *et al.*, "Mothers childhood hardship forecasts adverse pregnancy outcomes: Role of inflamma-

tory, lifestyle, and psychosocial pathways," *Brain, behavior, and immunity*, vol. 65, pp. 11–19, 2017.

[24] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," *arXiv preprint arXiv:1202.3775*, 2012.

[25] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, A. J. Smola *et al.*, "A kernel statistical test of independence," in *NIPS*, vol. 20, 2007, pp. 585–592.

[26] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *NIPS*, vol. 20, 2007, pp. 489–496.

[27] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2010.

[28] F. Rebelo, M. M. Schlüssel, J. S. Vaz, A. B. Franco-Sena, T. J. Pinto, F. I. Bastos, A. R. Adegboye, and G. Kac, "C-reactive protein and later preeclampsia: systematic review and meta-analysis taking into account the weight status," *Journal of hypertension*, vol. 31, no. 1, pp. 16–26, 2013.

[29] G. D. Ernst, L. L. de Jonge, A. Hofman, J. Lindemans, H. Russcher, E. A. Steegers, and V. W. Jaddoe, "C-reactive protein levels in early pregnancy, fetal growth patterns, and the risk for neonatal complications: the generation r study," *American journal of obstetrics and gynecology*, vol. 205, no. 2, pp. 132–e1, 2011.

[30] J. N. Felder, R. J. Baer, L. Rand, L. L. Jelliffe-Pawlowski, and A. A. Prather, "Sleep disorder diagnosis during pregnancy and risk of preterm birth," *Obstetrics & Gynecology*, vol. 130, no. 3, pp. 573–581, 2017.

[31] P. A. Braveman, K. Heck, S. Egerter, K. S. Marchi, T. P. Dominguez, C. Cubbin, K. Fingar, J. A. Pearson, and M. Curtis, "The role of socioeconomic factors in black–white disparities in preterm birth," *American journal of public health*, vol. 105, no. 4, pp. 694–702, 2015.

[32] B. McKinnon, S. Yang, M. S. Kramer, T. Bushnik, A. J. Sheppard, and J. S. Kaufman, "Comparison of black-white disparities in preterm birth between canada and the united states," *Canadian Medical Association Journal*, pp. cmaj–150 464, 2015.

[33] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A hilbert space embedding for distributions," in *International Conference on Algorithmic Learning Theory*. Springer, 2007, pp. 13–31.

[34] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of machine learning research*, vol. 3, no. Jul, pp. 1–48, 2002.

[35] C. R. Baker, "Joint measures and cross-covariance operators," *Transactions of the American Mathematical Society*, vol. 186, pp. 273–289, 1973.

[36] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyonds," *arXiv preprint arXiv:1605.09522*, 2016.

[37] R. D. Shah and J. Peters, "The hardness of conditional independence testing and the generalised covariance measure," *arXiv preprint arXiv:1804.07203*, 2018.

[38] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Transactions on Information Theory*, vol. 37, pp. 1501–1518, 1991.

[39] K.-C. Liang and X. Wang, "Gene regulatory network reconstruction using conditional mutual information," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2008, no. 1, p. 253894, 2008.

[40] A. J. Hartemink, "Reverse engineering gene regulatory networks," *Nature biotechnology*, vol. 23, no. 5, p. 554, 2005.

[41] X. Zhang, X.-M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.-K. Hao, Z.-P. Liu, and L. Chen, "Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information," *Bioinformatics*, vol. 28, no. 1, pp. 98–104, 2011.

[42] J. Numata, O. Ebenhöh, and E.-W. Knapp, "Measuring correlations in metabolomic networks with mutual information," in *Genome Informatics 2008: Genome Informatics Series Vol. 20.* World Scientific, 2008, pp. 112–122.

[43] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948, reprinted in [102].

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: John Wiley, 2006.

[45] R. M. Gray, *Entropy and Information Theory.* Springer Science & Business Media, 2011.

[46] A. Dembo, "Probability theory: Stat310/math230 apr 23, 2019," 2019.

[47] G. P. Basharin, "On a statistical estimate for the entropy of a sequence of independent random variables," *Theory of Probability & Its Applications*, vol. 4, no. 3, pp. 333–336, 1959.

[48] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[49] J. D. Victor, "Asymptotic bias in information estimates and the exponential (Bell) polynomials," *Neural Computation*, vol. 12, pp. 2797–2804, 2000.

[50] P. Grassberger, "Entropy estimates from insufficient samplings," *arXiv preprint physics/0307138*, 2003.

[51] E. Archer, I. M. Park, and J. W. Pillow, "Bayesian entropy estimation for countable discrete distributions," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2833–2868, 2014.

[52] I. Nemenman, W. Bialek, and R. D. R. Van Steveninck, "Entropy and information in neural spike trains: Progress on the sampling problem," *Physical Review E*, vol. 69, no. 5, p. 056111, 2004.

[53] A. Chao, Y. Wang, and L. Jost, "Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species," *Methods in Ecology and Evolution*, vol. 4, no. 11, pp. 1091–1100, 2013.

[54] Y. G. Dmitriev and F. Tarasenko, "On estimation of functionals of the probability density function and its derivatives," *Teoriya veroyatnostei i ee primeneniya*, vol. 18, no. 3, pp. 662–668, 1973.

[55] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[56] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[57] X. Wang, "Volumes of generalized unit balls," *Mathematics Magazine*, vol. 78, no. 5, pp. 390–395, 2005.

[58] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Transactions on Information Theory*, 2018.

[59] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[60] P. Zhao and L. Lai, "Analysis of knn information estimators for smooth distributions," *IEEE Transactions on Information Theory*, 2019.

[61] S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," *Physical review letters*, vol. 99, no. 20, p. 204101, 2007.

[62] M. Vejmelka and M. Paluš, "Inferring the directionality of coupling with conditional mutual information," *Physical Review E*, vol. 77, no. 2, p. 026214, 2008.

[63] A. Tsimpiris, I. Vlachos, and D. Kugiumtzis, "Nearest neighbor estimate of conditional mutual information in feature selection," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 697–12 708, 2012.

[64] J. Runge, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 938–947.

[65] A. Rahimzamani and S. Kannan, "Potential conditional mutual information: Estimators and properties," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2017, pp. 1228–1235.

[66] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Estimating mutual information for discrete-continuous mixtures," in *Advances in Neural Information Processing Systems*, 2017, pp. 5988–5999.

[67] A. Rahimzamani, H. Asnani, P. Viswanath, and S. Kannan, "Estimators for multivariate information measures in general probability spaces," in *Advances in Neural Information Processing Systems*, 2018, pp. 8664–8675.

[68] G. Casella and R. L. Berger, *Statistical inference*. Duxbury Pacific Grove, CA, 2002, vol. 2.

[69] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the compas recidivism algorithm," May 2016. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[70] Y. Hamamoto, S. Uchimura, and S. Tomita, "A bootstrap technique for nearest neighbor classifier design," *IEEE transactions on pattern analysis and Machine intelligence*, vol. 19, no. 1, pp. 73–79, 1997.

[71] S. Gao, G. Ver Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables," in *Artificial intelligence and statistics*, 2015, pp. 277–286.

[72] N. Carrara and K. Vanslette, "The design of global correlation quantifiers and continuous notions of statistical sufficiency," *Entropy*, vol. 22, no. 3, p. 357, 2020.

[73] K. R. Moon, M. Noshad, S. Y. Sekeh, and A. O. Hero, "Information theoretic structure learning with confidence," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 6095–6099.

[74] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[75] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[76] W. Rudin, *Real and complex analysis*. McGraw-Hill, 1987.

[77] G. P. Basharin, "On a statistical estimate for the entropy of a sequence of independent random variables," *Theory of Probability & Its Applications*, vol. 4, no. 3, pp. 333–336, 1959.

[78] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[79] J. D. Victor, "Asymptotic bias in information estimates and the exponential (bell) polynomials," *Neural computation*, vol. 12, no. 12, pp. 2797–2804, 2000.

[80] T. B. Berrett, R. J. Samworth, M. Yuan *et al.*, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *The Annals of Statistics*, vol. 47, no. 1, pp. 288–318, 2019.

[81] P. Zhao and L. Lai, "Analysis of knn information estimators for smooth distributions," *IEEE Transactions on Information Theory*, 2019.

[82] A. Hauser and P. Bühlmann, "Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2409–2464, 2012.

[83] D. Oehm, *Simulating data with Bayesian networks*, 2019. [Online]. Available: https://www.r-bloggers.com/simulating-data-with-bayesian-networks/

[84] G. Biau and L. Devroye, *Lectures on the nearest neighbor method*. Springer, 2015.

[85] R. E. Pérez-Figueroa, F. Kapadia, S. C. Barton, J. A. Eddy, and P. N. Halkitis, "Acceptability of PrEP uptake among racially/ethnically diverse young men

who have sex with men: The P18 study," *AIDS Education and Prevention*, vol. 27, no. 2, pp. 112–125, 2015.

[86] R. M. Grant, J. R. Lama, P. L. Anderson, V. McMahan, A. Y. Liu, L. Vargas, P. Goicochea, M. Casapía, J. V. Guanira-Carranza, M. E. Ramirez-Cardich *et al.*, "Preexposure chemoprophylaxis for hiv prevention in men who have sex with men," *New England Journal of Medicine*, vol. 363, no. 27, pp. 2587–2599, 2010.

[87] C. Hare, J. Coll, P. Ruane, J. Molina, K. Mayer, H. Jessen *et al.*, "The phase 3 DISCOVER Study: Daily F/TAF or F/TDF for HIV pre-exposure prophylaxis," in *Annual Conference on Retroviruses and Opportunistic Infections (CROI). Seattle*, 2019.

[88] D. K. Owens, K. W. Davidson, A. H. Krist, M. J. Barry, M. Cabana, A. B. Caughey, S. J. Curry, C. A. Doubeni, J. W. Epling, M. Kubik *et al.*, "Preexposure prophylaxis for the prevention of HIV infection: US Preventive Services Task Force recommendation statement," *JAMA*, vol. 321, no. 22, pp. 2203–2213, 2019.

[89] J. C. Hojilla, K. A. Koester, S. E. Cohen, S. Buchbinder, D. Ladzekpo, T. Matheson, and A. Y. Liu, "Sexual Behavior, Risk Compensation, and HIV Prevention Strategies among Participants in the San Francisco PrEP Demonstration Project: a qualitative analysis of counseling notes," *AIDS and Behavior*, vol. 20, no. 7, pp. 1461–1469, 2016.

[90] C. Grov, T. H. Whitfield, H. J. Rendina, A. Ventuneac, and J. T. Parsons, "Willingness to take PrEP and potential for risk compensation among highly sexually active gay and bisexual men," *AIDS and Behavior*, vol. 19, no. 12, pp. 2234–2244, 2015.

[91] M. E. Newcomb, K. Moran, B. A. Feinstein, E. Forscher, and B. Mustanski, "Pre-Exposure Prophylaxis (PrEP) Use and Condomless Anal Sex: Evidence of Risk Compensation in a Cohort of Young Men Who Have Sex with Men." *Journal of Acquired Immune Deficiency Syndromes (1999)*, vol. 77, no. 4, pp. 358–364, 2018.

[92] J. A. Schneider, A. Bouris, and D. K. Smith, "Race and the Public Health Impact Potential of Pre-Exposure Prophylaxis in the United States," *Journal of Acquired Immune Deficiency Syndromes*, vol. 70, no. 1, pp. e30–e32, 2015.

[93] M. Zlotorzynska, P. Sullivan, and T. Sanchez, "The Annual American Men's Internet Survey of Behaviors of Men who have Sex with Men in the United States: 2016 key indicators report," *JMIR Public Health and Surveillance*, vol. 5, no. 1, p. e11313, 2019.

[94] P. S. Sullivan, E. S. Rosenberg, T. H. Sanchez, C. F. Kelley, N. Luisi, H. L. Cooper, R. J. Diclemente, G. M. Wingood, P. M. Frew, L. F. Salazar *et al.*, ""explaining racial disparities in hiv incidence in black and white men who have sex with men in atlanta, ga: a prospective observational cohort study"," *Annals of Epidemiology*, vol. 25, no. 6, pp. 445–454, 2015.

[95] B. J. Hill, K. Rosentel, and L. Hebert, "Brief report: Assessing the impact of race on hiv/sti risk perceptions among young men who have sex with men using an experimental approach," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 81, no. 2, pp. 153–157, 2019.

[96] M. Holt, T. Lea, L. Mao, J. Kolstee, I. Zablotska, T. Duck, B. Allan, M. West, E. Lee, P. Hull *et al.*, "Community-level changes in condom use and uptake of hiv pre-exposure prophylaxis by gay and bisexual men in melbourne and sydney, australia: results of repeated behavioural surveillance in 2013–17," *The lancet HIV*, vol. 5, no. 8, pp. e448–e456, 2018.

[97] C. E. Oldenburg, A. S. Nunn, M. Montgomery, A. Almonte, L. Mena, R. R. Patel, K. H. Mayer, and P. A. Chan, "Behavioral changes following uptake of hiv pre-exposure prophylaxis among men who have sex with men in a clinical setting," *AIDS and Behavior*, vol. 22, no. 4, pp. 1075–1079, 2018.

[98] O. **Mesner**, A. Davis, E. Casman, H. Simhan, C. Shalizi, L. Keenan-Devlin, A. Borders, and T. Krishnamurti, "Using Graph Learning to Understand Adverse Pregnancy Outcomes and Stress Pathways," *PloS One*, vol. 14, no. 9, p. e0223319, 2019.

[99] O. C. **Mesner** and C. R. Shalizi, "Conditional Mutual Information Estimation for Mixed Discrete and Continuous Variables with Nearest Neighbors," *arXiv preprint arXiv:1912.03387*, 2019.

[100] M. Raginsky, "Directed information and Pearl's causal calculus," in *Proceedings of the 49th Annual Allerton Conference on Communication, Control and Computing*, S. Meyn and B. Hajek, Eds. IEEE, 2011, pp. 958–965. [Online]. Available: http://arxiv.org/abs/1110.0718

[101] I. Kontoyiannis and M. Skoularidou, "Estimating the directed information and testing for causality," *IEEE Transactions on Information Theory*, vol. 62, pp. 6053–6067, 2016. [Online]. Available: https://arxiv.org/abs/1507.01234

[102] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press, 1963.