Matrix-Variate Graphical Models for High-Dimensional Neural Recordings

Zongge Liu

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Robert E. Kass, Chair
Zhao Ren
Valerie Venture
Cosma Shalizi

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Keywords: Matrix-variate Gaussian Graphical Model, Simultaneous Testing, Multi-graph Estimation, Precision Matrix Estimation

For my family

Abstract

As large-scale neural recordings become common, many neuroscientific investigations are focused on identifying functional connectivity from spatio-temporal measurements in two or more brain areas. Spatio-temporal data in neural recordings can be viewed as matrix-variate data, where the first dimension is time and the second dimension is space. A matrix-variate Gaussian Graphical model (MGGM) can be applied to study the conditional dependence between different nodes under the assumption that the overall covariance is the tensor product of the spatial covariance and temporal covariance. This provides a way to study functional connectivity via the spatial component of the precision matrix. We develop and study penalized regression methods that enable us to do statistical inference and simultaneous hypothesis testing for multi-session local field potential data. Our approach includes four innovations. First, we provide a simultaneous testing framework for MGGM with a high-dimensional bootstrap technique, which enables us to test the strength of neural connectivity between two brain areas. Second, because estimation of spatial dependence also relies on an accurate estimate for temporal covariance structure, we assume autoregressive temporal dependence and thereby provide estimation of the temporal precision matrix based on a Cholesky factor decomposition. Third, for spatial precision matrix estimation and inference, we implement group Lasso to jointly estimate multi-graphs and study a new statistic to aggregate information from multiple sessions to improve inference. Finally, using our matrix-variate assumption for high-dimensional data, we develop a novel cross-region dynamic factor analysis model to estimate dynamic neural connectivity across multiple brain regions.

Acknowledgments

I'm very grateful for these wonderful years at Carnegie Mellon University, working with so many kind and brilliant people. I'm extremely fortunate to work with my thesis advisor, Prof. Robert E. Kass, and external advisor, Prof. Zhao Ren. I'm also thankful for the support from my family and friends.

Contents

List of Figures

List of Tables

Chapter 1

Background

Neurons communicate with each other via action potentials, which are rapid electrical events and discharges. As large neural recordings become common, many neuroscientific investigations are focused on identifying associations from spatio-temporal measurements within one area and across different brain areas [62]. These studies attempt to answer the question whether the neural activity changes across different time stages and experimental conditions. Local Field Potentials (LFPs) are electrical signals generated by the synchronous electrical activity of the neurons near the electrodes, and LFP is believed to be correlated with subthreshold membrane potential fluctuations [45][11][29]. Past work reveals that LFP is also highly correlated with network fluctuations in single neuron spiking activity, and decodes spike–stimulus relationship [33]. As recent progress in technology enables neural measurements with finer resolution and larger volumes [56], it also poses new challenges to develop tools to analyze the data with high-dimensional statistics. LFP data usually consists of multi-electrode measurements spanning over a certain time range. For example, we have a three-dimensional dataset with 3000 trials, 192 electrodes and 800 timeticks (800ms). The problem is to characterize statistical associations which we call connectivity, examine and test changes in connectivity within brain areas and across brain areas. To make it even more challenging, the same experiment may be repeated over multiple sessions (days) for one subject, which we call the multi-session data. To solve this problem, we analyze spatial-temporal data under Matrix-variate Gaussian Graphical Model, from both inference-focused side and estimation-focused side. For inference part, we develop a Matrix-variate Gaussian Graphical Model and fit it using the regression-based approach [38][51], and provide a simultaneous testing based on Gaussian approximation bootstrap method proposed by [17]. We further formulated the multi-session data inference into a multi-graph inference problem, and provide a novel estimator for multiple Matrix-variate Gaussian Graphical Model. For estimation part, we propose a novel Latent Dynamic Factor Analysis for High-dimensional (LDFA-H) data combining flavors of both canonical correlation analysis and factor analysis to improve the estimation of cross-region dynamic connectivity, which not only recovers dynamic cross-region connectivity, but also retrieves low-dimensional temporal factors and spatial factor loadings for neural population.

1.1 Experiment

In our thesis, we analyze data from multiple LFP recordings in visual cortex area V4 (V4) and prefrontal cortex (PFC) which are believed to involved in visual processing. V4 is a mid-level visual region in visual-processing where inputs from earlier regions are further processed and transmitted to later regions. Effects of attention on neuronal responses of visual cortex are highlighted by several electrophysiological studies [46][4]. Moreover, PFC has long been considered to be involved with control of attention and response modulation in posterior visual areas. The dataset was collected by the Smith Lab at Carnegie Mellon Neuroscience Institute during an experiment on the neurophysiology of attention [32]. During the experiment, a monkey is asked to perform a memory-guided saccade task. The timeline of the experiment is stated as follows:

- The animal fixated on a point at the center of the screen for 200ms.
- A circular target appeared at one of eight randomly chose locations of the screen for 50ms.
- The animal had to remember the location of the target while maintaining fixation for a delay period of 500ms.
- After the delay period, the fixation point was turned off, and the monkey had to make a saccade to remember the location of the target.

Local field potentials were measured simultaneously from multiple electrodes in PFC and V4 while monkey was performing the task. After data preprocessing, only successful trials are kept in the dataset. There are 2000 trials, 192 electrodes (96 in PFC and 96 in V4) and 750 timeticks (750ms long) in total. There is also the location information for the electrodes in each area. A subset of the data containing the first trial is shown in Figure 1.1. The measurement covered all experimental stages described as above and an additional 50ms after the delay period. The color in the heatmap corresponds to the intensity of LFP signal, and V4 seems to be more active than PFC.

The same experiment session might be repeated by multiple times as well, which results in multi-session LFP recordings. In this thesis, we obtain multi-session dataset from the Smith Lab, where during each session the animal is performing the same saccade task as before. The complete dataset contains 5 consecutive sessions, 2000 trials, 192 electrodes and 750 timeticks.

1.2 Data Analysis Goals

For this dataset, we are interested in the following questions:

- **Q1.** How should we characterize the connectivity among large number of LFPs while taking advantage of their spatio-temporal structures?
- **Q2.** How strong is the connectivity among electrodes within PFC and V4? How does connectivity change when the animal is under different conditions?
- **Q3.** How strong is the connectivity across these two areas?

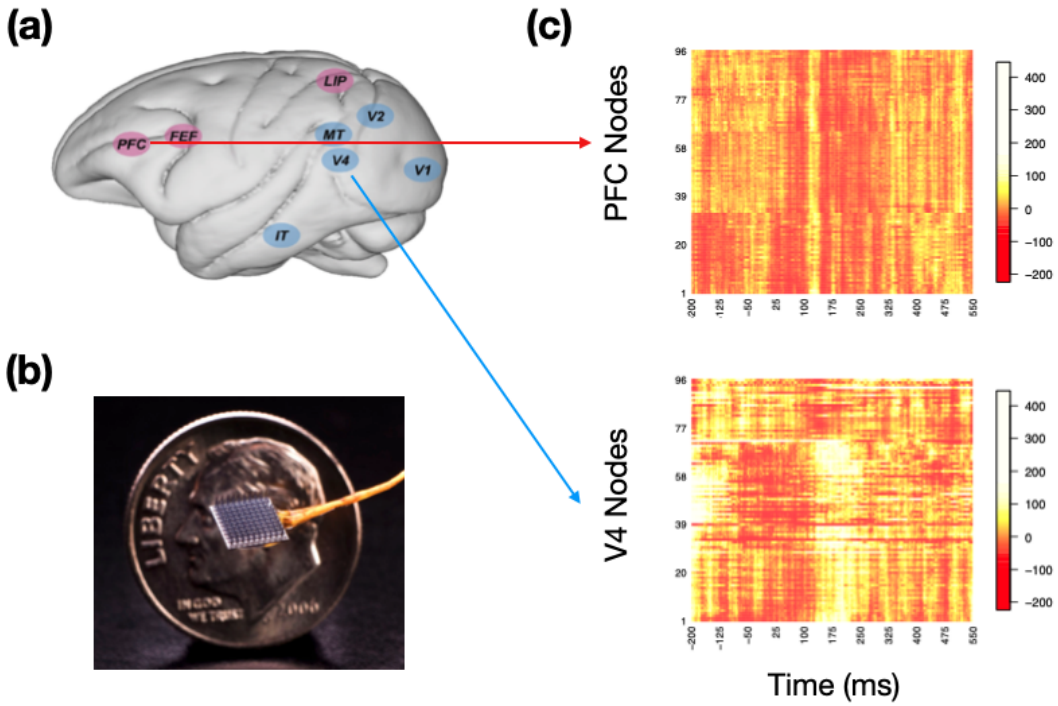

Figure 1.1: Experimental data from monkey saccade task. (a) Primate cortical areas of the attention network [53]. Pink areas are the approximate locations of attention control areas while blue areas are the approximate locations of visual areas. (b) Utah array with 10x10 recording electrodes with $400\mu m$ interval [43]. For each region, one Utah array with 96 electrodes is implanted. (c) LFPs were recorded simultaneously from V4 and PFC. The X axis is time in millisecond, while Y is the electrode. Time is aligned at $t = 0$ for each trial when the circular target just appeared. The total length is 750ms covering all experimental stages.

1.3 Gaussian Graphical Model

The vector-variate Gaussian Graphical model has been widely applied to high dimensional data in scientific studies to explore the conditional dependence relationship among entries of a random vector, including the neural data [25, 60]. For a random vector $\boldsymbol{Z} \in \mathbb{R}^p$, we can define a undirected graph $G = (V, E)$ associated with $\boldsymbol{Z}$, where the set V contains p nodes each of which corresponds to an entry in $\boldsymbol{Z}$, and the set E consists of all edges among V. Specifically, there is no edge between two nodes in V if and only if the corresponding two variables are conditionally independent given the rest of variables in $\boldsymbol{Z}$. For a Gaussian vector, one can assess the graph structure, i.e., the set E, in terms of the precision matrix, the inverse of covariance matrix of $\boldsymbol{Z}$ [34] because two variables are conditionally independent if and only if their partial correlation is zero. Now consider that we have n samples $\{\boldsymbol{x}^{(i)}\}_{i=1}^n$ where $\boldsymbol{x}^{(i)} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$; the log-likelihood can be written as

$$l(\boldsymbol{\Omega}) = \frac{1}{2}\log\det(\boldsymbol{\Omega}) - \frac{1}{2}\text{tr}(\boldsymbol{S\Omega}), \tag{1.1}$$

where the sample covariance matrix $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^{(i)}(\boldsymbol{x}^{(i)})'$.

To answer substantive scientific questions **Q2**-**Q3**, estimation and inference for the entry in the precision matrix are required. Methods to estimate the sparse precision matrix for normal distribution are well-studied in recent years [44][66]. There are two main categories of method for sparse precision matrix estimation: penalized likelihood method [1] and regression based method [38][51].

The penalized likelihood method is a popular method for Gaussian Graphical Model estimation and inference. This method adds a penalty term in addition to Equation 1.1. For example, in Graphical Lasso [26], the goal is to maximize the following likelihood function

$$l(\boldsymbol{\Omega}) = \log\det(\boldsymbol{\Omega}) - \text{tr}(\boldsymbol{S\Omega}) + \rho\|\boldsymbol{\Omega}\|_1,$$

where ρ is the tuning parameter. For statistical inference, [31] proposed a method to estimate confidence band for entries in sparse precision matrices, which guaranteed asymptotic normality for each edge. However, since the objective function is non-convex, the optimization approach is theoretically less appealing; moreover, methods based on optimization will usually require the irrepresentability condition, which is difficult to validate. We adapted the regression-based approach in our work [38][51]. Assuming $1 \leq i \leq p$, $1 \leq j \leq p$, we consider the regression model as follows,

$$X_i = \boldsymbol{X}_{-i}\boldsymbol{\beta}_i + \varepsilon_i = \sum_{j \neq i} \boldsymbol{X}_j \beta_{i,j} + \varepsilon_i,$$

where $\boldsymbol{\beta}_{i,j} = -\frac{\Omega_{ij}}{\Omega_{ii}}$. It follows from the linear regression theory that $\text{Cov}(\varepsilon_i, \varepsilon_j) = \frac{\Omega_{ij}}{\Omega_{ii}\Omega_{jj}}$. Therefore, we can estimate the entries in precision matrix with the regression residuals. Following this regression approach, the work of [49] studied the estimation of c-level partial correlation graphs in ordinary graphical models where partial correlations are larger than a pre-specified constant c. Moreover, the work of [15] studied the construction for simultaneous confidence regions for the precision matrix in a Gaussian graphical model.

Because LFP data involve both space (electrodes on an array) and time, we replace the Gaussian graphical model by a Matrix-variate Graphical Model (MGGM). Here, the recordings from each trial form a $p \times q$ matrix, where p is the number of time points and q is the number of electrodes. MGGM has become popular in analyzing spatio-temporal data from biomedical imaging and financial markets [68][16].

1.4 Matrix-variate Graphical Model

We let $\text{vec}(\boldsymbol{X}) \in \mathbb{R}^{pq\times 1}$ denote the vectorization of matrix $\boldsymbol{X}$, $\otimes$ as the Kronecker product. It follows that

$$\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{U} \otimes \boldsymbol{V})$$

if and only if

$$\text{vec}(\boldsymbol{X}^{'}) \sim N(\text{vec}(\boldsymbol{\mu}^{'}), \boldsymbol{U} \otimes \boldsymbol{V}), \tag{1.2}$$

with mean $\boldsymbol{\mu} \in \mathbb{R}^{p\times q}$, row covariance matrix $\boldsymbol{U} \in \mathbb{R}^{p\times p}$, and column covariance matrix $\boldsymbol{V} \in \mathbb{R}^{q\times q}$. By fitting our data with MGGM, we assume that the trials are repeated under the same condition, the LFP temporal correlation is consistent across space and the spatial correlation is consistent across time.

We can design our hypothesis testing for each edge (**Q2**) by deriving the corresponding asymptotic distribution for the entry in spatial precision matrix. Testing a set of edges is more theoretically challenging. However, there are many pioneer works focusing on the graph estimation [37][64][67][68], and interesting recent works on inference for Matrix-variate Gaussian graphical model using the regression based approach.

The work of [16] developed a multiple testing framework for support recovery in MGGM, and provided theoretical analysis for asymptotic normality and false discovery rate (FDR) control. The work of [63] developed a paired test of matrix graphs to infer brain connectivity with correlated samples, and similarly, their testing procedure was based on multiple testing and FDR control. We are interested, however, in deriving a single statistic to simultaneously test the conditional dependence between subgroups of nodes.

The most relevant work is [61]. The authors proposing a multiple testing framework to identify conditionally independent pairs, and developed another global test to examine whether all spatial locations are conditionally dependent. The major limitation is that the limiting distribution of test statistic is Gumbel distribution, which has a notoriously slow rate of convergence approximation [39]. Besides, the global test included the tests of all off-diagonal edges in the graph, while in our design the subset of edges can be of arbitrary size. Finally, most of the previous works will either assume that the temporal covariance is known, or only estimate them with sparsity assumption, while neglecting the decaying autocorrelation structure of the temporal signals.

Therefore, to handle **Q2**, we formulate the simultaneous testing procedure using Gaussian approximation bootstrap method proposed by [17]. We can test the strength of neural connectivity between two brain areas by testing the corresponding subset of edges in precision matrix. With testing a single edge and testing a group of edges solved under MGGM model, we will be able to identify the change of connectivity by deriving the asymptotic distribution of the difference between test statistics. We noticed that previous methods analyzing LFP signals such as Granger Causality model the temporal data as an vector auto-regressive model [55], and accordingly, we propose a Cholesky factor decomposition approach [5] which imposes a non-stationary condition for temporal precision matrix estimation, while still assuming sparsity for spatial precision matrix estimation.

For multi-session data, each session can be viewed as a single matrix-variate graph. However, naively applying the method developed for single graph to multi-session data may lose information about the common structures shared across different graphs. Especially, since the task and the subject are the same across all sessions, we can assume that the structure of the connectivity (the sign of edge between two particular electrodes in spatial precision matrices) remains the same, while the amplitude might vary across different graphs. Therefore, to fully address **Q2** in such dataset, a novel multiple MGGM estimation and inference method is needed.

Inference under multiple matrix-variate graphs is missing in literature. The most relevant work is by [68], where the main focus is on multiple matrix-variate graph estimation. The authors established a non-convex optimization method with sparse and group Lasso penalization to estimate multiple matrix-variate Gaussian graphs for matrix-normal distributed data. They further designed an efficient optimization algorithm, and established the asymptotic properties of the estimator under special scenarios with sparse penalty or group penalty only. However, since this work was based on non-convex optimization method, stronger assumptions are needed to guarantee convergence of estimation, and inference remains to be unknown under such a

formulation. Besides, estimation of multiple ordinary Gaussian graphical models based on optimization method was studied by [12, 20, 36]. On the other hand, [48] proposed Bayesian inference on multiple Gaussian graphical models. A Markov random field prior is implemented to encourages common edges across graphs, and a spike-and-slab prior is placed on the parameters to learn the groups which have a shared structure. Therefore, this model can learn the information between sample groups, and measure the relative network similarity across groups effectively. In contrast, our new inference framework is motivated by inference with regression based method in multiple ordinary networks. [50] proposed a large-scale tuning-free heterogeneous inference framework with on chi-based and linear functional-based tests under ordinary Gaussian graphical models. Especially, the linear functional-based test is optimal in achieving testable region boundary, and the sample size requirement for the linear functional-based test is minimal. With regression method based on Group Lasso, we adapt linear functional-based test into Matrix-variate graph setting, and extend the simultaneous testing under single-graph case to multiple-graph case. We show that in both theory and simulation, our method outperforms baseline methods [12, 36, 50, 68] with better estimation accuracy and test power.

However, when moving from **Q2** to **Q3** and considering cross-region inference, one single matrix-variate graphical model applying to all regions may be too stringent. The Kronecker product assumption indicates that all electrodes share the same autocorrelation structure, which barely holds for multiple brain regions. Therefore, to solve **Q3**, we move on with a more complicated factor model which can capture cross-region dynamic connectivity.

1.5 Latent Dynamic Factor Analysis of High-Dimensional Neural Recordings

Estimating dynamic connectivity structure across regions, which is a major concern of **Q3**, can be challenging due to the high dimensionality in spatial and temporal domain. Moreover, the auto-correlation of noise in the measurements is usually strong, which may contaminate the cross-region connectivity structure and make it difficult to detect. One simple yet popular approach is a two-step approach, where researchers take averages over signals in each area and apply Granger causality [55] to study information flow between regions[10]. However, taking simple averages means losing information from each unique signal inside one area, especially considering the scenario where a subset of signals in one region are positively correlated with those in the other region, while the correlation between another subset of signals might be negative. In recent years, people developed new method based on canonical correlation analysis (CCA) to handle dynamic connectivity. Dynamic kernel-CCA (DKCCA) extends traditional CCA by implementing a more flexible and computationally efficient kernel-based approach and allowing the correlation structure dynamically changing over time [52]. Compared with other common cross-correlation analysis method, DKCCA discovers a dynamic switch in the lead-lag relationship between the hippocampus (HPC) and prefrontal cortex (PFC). [7] extended a parametric approach on CCA to develop a model-based cross-region connectivity estimate DynaMICCS. Adopting a Gaussian graphical model, the authors recapitulated cross-region connectivity in terms of partial-correlation graphs. The most serious issue with CCA-based approach is that it fails to account

for the contribution of the correlation structure from auto-correlated noise inside each region. Without properly denoising the signal and smoothing the data, it is questionable whether the cross-connectivity is contaminated or even obscured by the noise.

On the other hand, to study cross-connectivity, factor analysis model, and its dynamical generalization, state space model draw more and more attention in neuroscience community over recent years. Such models are shown to provide explainable low-dimensional representation, and generate better characterization of the population activity than traditional two-stage approaches. Gaussian Process Factor Analysis (GPFA) offers a flexible and stable framework for extracting smooth low-dimensional latent factors [65]. Despite its popularity in neuroscience community, GPFA fails to address the cross-region connectivity and lead-lag issues, thus becomes inappropriate for analyzing large neural recordings containing measurements from multiple brain regions. On the other hand, a state space model was proposed to tackle with cross-region dynamic connectivity in magnetoencephalography (MEG) and electroencephalography (EEG) [62]. This model characterizes the non-stationary dynamic dependence across regions of interest (ROIs) using time-varying auto-regression, and the mean response of each ROI is encoded as the space variable. While it addresses the cross-region dynamic connectivity to some extent, it remains to have several serious issues. First, both auto-regressive assumption for factors and independent assumption for noises appear to be rigid and limited, which barely hold under any realistic scenario. Second, the model does not include sparsity-inducing regularization, therefore the number of parameters can be large and the model can be under-determined under high-dimensional settings.

In this thesis, we propose a new model which uses probabilistic CCA to carry out an extension of factor models, but the framework allows far richer spatiotemporal dependencies than is typically assumed in GPFA, and it has been built to handle high-dimensional problems. Here, "spatial" dependence refers to dependence among the various observational time series and, in the neural context, this results from the spatial arrangement of the electrodes, each of which records one of the time series. In the usual setup of GPFA, two reasonable simplifications are made: first, the observation noise is assumed to be white and, second, the latent Gaussian processes are stationary. Our approach relaxes these assumptions: We allow the observation noise to have spatiotemporal dependence, which we have found more realistic, and we let the latent processes be non-stationary, so their dependence can evolve dynamically, which is important in our applications because cross-process dependence describes the sudden flow of information from one brain region to another during short epochs. We thus call our method LDFA-H. These generalizations come at a cost: we now have a high-dimensional time series problem within each brain region together with a high-dimensional covariance structure. We solve these high-dimensional problems by imposing sparsity of the dominant effects, building on related, but as yet unpublished work [7] that treats the high-dimensional covariance structure in the context of observational white noise, and by incorporating banded covariance structure as in [5]. In a simulation study, based on realistic synthetic time series, we verify recovery of cross-region structure even when some of our assumptions are violated, and even in the presence of high noise. We then apply the method to our experimental data and find time-varying cross-region dependencies.

1.6 Organization of this Document

My thesis focus is on developing new inference and estimation method for spatial-temporal neural data under the framework of matrix-variate Gaussian graphical models to study the neural connectivity within area and cross area. In Chapter 2, I provide a procedure for multiple matrix-variate graph estimation and inference based on regression method, which effectively addresses **Q1**-**Q2**. In Chapter 3, for cross-area dynamic connectivity analysis, an estimation-focused factor analysis model is designed to answer **Q3**.

Chapter 2

Inference and Estimation in Multiple Matrix-Variate Graphical Models

This chapter is taken from work submitted to *Journal of the American Statistical Association in 2020* aside from minor changes for style consistency. I collaborated with Zhao Ren and Robert E. Kass.

The primary goal of this chapter is to address **Q2** – how strong the connectivity and change of connectivity are within each brain area. We fit the spatio-temporal LFPs in each session with a Matrix-variate Gaussian Graphical Model, as the spatial connectivity can be characterized by spatial precision matrix. The multiple Matrix-variate Graphs are then estimated using Group Lasso and we developed a new statistic which aggregates information across all sessions. Based on the test statistic, the hypothesis testing framework for both single edge case and multiple edge case are studied under multiple MGGMs. We further formulate **Q2** into three types of hypothesis tests as in Equation (2.1)-(2.3).

We consider d matrix-variate Gaussian graphs which encode neural connectivity patterns among p time ticks and q spatial locations during d different experimental sessions. For each dataset on session t, we have n independent and identically distributed (i.i.d) $p \times q$ matrix-variate samples, $X_t^{(1)}$,...,$X_t^{(n_t)}$, following matrix-variate Gaussian distribution which is defined in Equation (1.2). For any $t \in [d]$ and $i \in [n_t]$,

$$\boldsymbol{X}_t^{(i)} \sim N(\mathbf{0}, \boldsymbol{U}_t \otimes \boldsymbol{V}_t).$$

Furthermore, let $\boldsymbol{A}_t = \boldsymbol{U}_t^{-1}$ denote the row precision matrix, and $\boldsymbol{B}_t = \boldsymbol{V}_t^{-1}$ denote column precision matrix. It is easy to see that $\boldsymbol{X}_t^{(i)} \sim N(\mathbf{0}, (\boldsymbol{A}_t \otimes \boldsymbol{B}_t)^{-1})$, which implies that the graph structure is encoded by $\boldsymbol{A}_t \otimes \boldsymbol{B}_t$. In particular, the spatial connectivity structure is the support of the column precision matrix $\boldsymbol{B}_t$. Comparing with precision matrix, partial correlation is preferred to characterize the magnitude of neural connectivity since it is invariant to scaling of variables. For column precision matrix, we define the corresponding partial correlation for each entry (i, j) at session t as $\rho_{tij} = -\frac{b_{tij}}{(b_{tii}b_{tjj})^{1/2}}$. We define $\mathcal{D} = \{\{\boldsymbol{X}_1^{(i)}\}_{i=1}^{n_1}, ..., \{\boldsymbol{X}_d^{(i)}\}_{i=1}^{n_d}\}$ as the collection of the full observations over these sessions.

In neural data, as is shown later in our data analysis (Figure 2.5), usually all sessions share a similar neural connectivity structure but the distributions of edge for each session may not be the same. Therefore, it is reasonable to assume $\boldsymbol{B}_t$'s may be different but share a similar support for $1 \leq t \leq d$. Equivalently, denoting the joint spatial partial correlation vector of the pair (i, j) by $\boldsymbol{\rho}_{ij}^0 = (\rho_{1ij}, \rho_{2ij}, \cdots, \rho_{dij})'$, we expect that either $\boldsymbol{\rho}_{ij}^0 = \mathbf{0}$ or it can be different from zero with distinct nonzero entries. We are commonly interested in testing whether the connectivity between a fixed pair of nodes or different regions exists across multiple sessions. In the single-edge test scenario, for a pair (i, j), we are interested in testing the following null hypothesis,

$$H_{0,ij} : \boldsymbol{\rho}_{ij}^0 = \mathbf{0}. \tag{2.1}$$

In the multiple edge test scenario, for a cross-region set $\mathcal{S}$, we aim to test whether there is no edge at all in $\mathcal{S}$, which is stated as the following null hypothesis,

$$H_{0,\mathcal{S}} : \boldsymbol{\rho}_{ij}^0 = \mathbf{0}, \ \forall (i, j) \in \mathcal{S}. \tag{2.2}$$

Notice that the single edge test can be treated as a special case of multiple edge test when $\mathcal{S} = \{(i, j)\}$. Moreover, we notice that partial correlation is considered to be closely related to effective connectivity and can be viewed as a graphical representation for interactions in biological system, such as neurons and genes [23, 41, 49]. Therefore, to study the strength of connectivity, a more interesting and practical hypothesis test is proposed,

$$H'_{0,\mathcal{S}} : \frac{1}{d} \sum_{t=1}^{d} |\rho_{tij}| \leq c, \ \forall (i, j) \in \mathcal{S}, \tag{2.3}$$

To address the above hypothesis tests effectively, our approach is based on the observation that using the d-session data collectively can achieve better accuracy than estimating each graph separately and combining the results naively. In this spirit, we assume that spatial precision matrices share similar sparsity structure and temporal precision matrices are banded following similar parametric space. Especially, for the spatial precision matrix, we also assume that the sign of spatial connectivity does not change across d sessions, which is reasonable since the locations of spatial nodes are fixed and the tasks the animal performs in each session are the same. The precise definition can be found in Section 2.2.2.

Our new inference framework is motivated by the following works targeting to three different aspects of our main goal. For testing the strength of conditional dependence, the work of [49] studied the estimation of c-level partial correlation graphs in ordinary graphical models where partial correlations are larger than a pre-specified constant c. For inference in multiple networks, [50] proposed a large-scale tuning-free heterogeneous inference framework with on chi-based and linear functional-based tests under ordinary Gaussian graphical models. Especially, the linear functional-based test is optimal in achieving testable region boundary, and the sample size requirement for the linear functional-based test is minimal. For simultaneous testing under MGGM, the most relevant work is [61]. The authors proposed a multiple testing framework to identify conditionally independent pairs, and developed another global test to examine whether all spatial locations are conditionally dependent. The major limitation is that the limiting distribution of test statistic is Gumbel distribution, which has a notoriously slow rate of convergence approximation

[39]. Besides, the global test includes the tests of all off-diagonal edges in the graph, while in our design the subset of edges can be of arbitrary size.

In our work, we propose a three-step procedure targeting to the hypothesis tests in Equation (2.1)-(2.3). First, we estimate our spatial correlation matrix using residuals and coefficients from group Lasso using data from all sessions, and reconstruct the spatial precision matrix and partial correlation matrix. Second, we follow the modified Cholesky factor decomposition to estimate the Frobenius norm of temporal covariance matrix for each graph separately. Finally, based on the results from the previous two steps, we construct a new linear-functional test statistic, which effectively address both single edge test and multiple edge test. The main contributions of the paper are summarized as follows:

- For spatial partial correlation, by assuming the sign of connectivity invariant across multiple graphs, we establish the theoretical guarantee of group Lasso to estimate multiple matrix-variate graphs with high-dimensional multi-response regression. Such a joint-estimation approach enjoys more efficiency and fewer samples when comparing with naive approach which applies single matrix-variate graph method on each graph.
- For temporal covariance, most of the previous works will either assume that the temporal covariance is known, or only estimate them with sparsity assumption, while neglecting the decaying autocorrelation structure of the temporal signals. We noticed that previous methods analyzing LFP signals such as Granger Causality fit the temporal data with an vector auto-regressive model [55], and accordingly, we propose a Cholesky factor decomposition approach [5] which imposes a non-stationary condition for temporal precision matrix estimation.
- We formulate the simultaneous testing procedure using Gaussian approximation bootstrap method proposed by [17]. This framework enables us to test the strength of neural connectivity between two brain areas by testing the corresponding subset of edges over multiple partial correlation matrices simultaneously, thus address the issue of simultaneous test in multiple MGGMs for the first time.

2.1 Simultaneous Inference in Multiple Matrix-variate Networks

In this section, we develop our inference framework aiming to hypothesis testing problems stated in (2.1)-(2.3), and leave all theoretical studies in Section 2.3. The full procedure is summarized in Algorithm 1 at the end of this section. For each dataset on session $1 \leq t \leq d$, we have n_t i.i.d $p \times q$ matrix-variate samples, $\boldsymbol{X}_t^{(1)}, ..., \boldsymbol{X}_t^{(n_t)}$, following matrix-variate Gaussian distribution $N(\boldsymbol{0}, \boldsymbol{U}_t \otimes \boldsymbol{V}_t)$, where n_t can vary among different sessions. We further define $\mathcal{D} = \{\{\boldsymbol{X}_1^{(i)}\}_{i=1}^{n_1}, ..., \{\boldsymbol{X}_d^{(i)}\}_{i=1}^{n_d}\}$ as the collection of the full observations over these sessions.

2.1.1 Notations and Preliminaries

We use bold notation for all matrices and vectors. We adopt the following notation throughout this paper. For a p-dimensional vector $\boldsymbol{x}$, we write $\boldsymbol{x}_{i:j}$ for its sub-vector $(x_i, x_{i+1}, \cdots, x_j)'$. Moreover,

let $\|\boldsymbol{x}\|_p$ denote the vector ℓ_p-norm of $\boldsymbol{x}$, and $J(\boldsymbol{x}) = \{1 \leq k \leq p : x_k \neq 0\}$. Assume that $\boldsymbol{X}$ is a $p \times q$ matrix, let $\boldsymbol{X}_{i,\cdot}$ denote the i-th row of $\boldsymbol{X}$, $\boldsymbol{X}_{\cdot,j}$ denote the j-th column of $\boldsymbol{X}$, $\boldsymbol{X}_{i,-j}$ denote the i-th row of $\boldsymbol{X}$ with j-th column removed, $\boldsymbol{X}_{-i,j}$ denote the j-th column of $\boldsymbol{X}$ with i-th row removed, and $\boldsymbol{X}_{-i}$ denote the original matrix with i-th row and i-th column removed. Let $\|\boldsymbol{X}\|_p$ denote the matrix p-norm of $\boldsymbol{X}$, while notice that $\|\boldsymbol{X}\|_2$ is also the spectral norm or operator norm. We use $\|\boldsymbol{X}\|_F$ to denote its Frobenius norm and $|\boldsymbol{X}|_\infty$ to denote the entry-wise sup-norm. For any set J we denote its cardinality by $|J|$. Finally, for the spatial graphs, we denote the non-zero spatial edge set in the graphs by $\mathcal{E}_s = \big\{(i,j) : 1 \leq i \neq j \leq q \text{ and } \boldsymbol{\rho}_{ij}^0 \neq \mathbf{0}\big\}$.

2.2 Simultaneous Inference Framework

In this section, we develop our inference framework aiming to hypothesis testing problems stated in (2.1)-(2.3), and leave all theoretical studies in Section 2.3. The full procedure is summarized in Algorithm 1 at the end of this section. For each dataset on session $1 \leq t \leq d$, we have n_t i.i.d $p \times q$ matrix-variate samples, $\boldsymbol{X}_t^{(1)}, ..., \boldsymbol{X}_t^{(n_t)}$, following matrix-variate Gaussian distribution $N(\mathbf{0}, \boldsymbol{U}_t \otimes \boldsymbol{V}_t)$, where n_t can vary among different sessions. We further define $\mathcal{D} = \{\{\boldsymbol{X}_1^{(i)}\}_{i=1}^{n_1}, ..., \{\boldsymbol{X}_d^{(i)}\}_{i=1}^{n_d}\}$ as the collection of the full observations over these sessions.

2.2.1 Estimation of Spatial Precision Matrix

We first discuss the model $\boldsymbol{X}_t \sim N(\mathbf{0}, (\boldsymbol{A}_t \otimes \boldsymbol{B}_t)^{-1})$ for each graph/session $1 \leq t \leq d$ as a motivation for our approach. Node-wise regression ([38, 44]) has been a popular approach for graphical model analysis. Indeed, it is well-known that the conditional distribution of X_{tli} (i.e., electrode i at time l) against remaining variables $\boldsymbol{X}_{tl,-i}$ at the same time l follows a linear regression,

$$X_{tli} = \boldsymbol{X}_{tl,-i}\boldsymbol{\beta}_{ti} + \varepsilon_{tli} = \sum_{j \neq i} \boldsymbol{X}_{tlj}\beta_{tij} + \varepsilon_{tli},$$

where it follows from the linear regression theory that $\beta_{tij} = -\frac{b_{tij}}{b_{tii}}$ and $\mathbb{E}(\varepsilon_{tli}) = 0$. Due to this connection between the coefficient $\boldsymbol{\beta}_{ti}$ and the column precision matrix $\boldsymbol{B}_t$, a certain sparsity assumption on $\boldsymbol{B}_t$ naturally implies a sparse linear regression model. On the one hand, given the values of $\boldsymbol{\beta}_{ti}$ for each time $1 \leq l \leq p$, it is easy to obtain that the covariance matrix among all residuals across $1 \leq i \leq q$, i.e., $\boldsymbol{\varepsilon_{tl}} := (\varepsilon_{tl1}, ..., \varepsilon_{tlp})'$ is $\boldsymbol{R}_t^l = \mathrm{Cov}(\boldsymbol{\varepsilon_{tl}}) = (r_{tij}^l)_{q\times q} = (\frac{u_{tll}b_{tij}}{b_{tii}b_{tjj}})_{q\times q}$. Therefore, testing whether $\rho_{tij} = 0$ (or $\beta_{tij} = 0$) is equivalent to testing whether $r_{tij}^l = 0$. On the other hand, we note that the coefficient $\boldsymbol{\beta}_{ti}$ remains the same across all $1 \leq l \leq p$. In addition, $\rho_{tij} = 0$ if and only if all $r_{tij}^l = 0$ at all time tick $1 \leq l \leq p$. These two facts together suggest us to treat each row $\boldsymbol{X}_{tl,\cdot}$ of the matrix-variate $\boldsymbol{X}_t$ as a q-dimensional sample and consider all p samples together. By doing this, we have p correlated vector-variate samples for a sparse linear regression model, where the covariance among these "row samples" is characterized by $\boldsymbol{U}_t$. Finally, we define the average of the covariance matrices of residuals from these correlated p samples as $\boldsymbol{R}_t = (r_{tij})_{q\times q} = \frac{1}{p}\sum_{l=1}^p \boldsymbol{R}_t^l$ with

$$r_{tij} = \frac{\mathrm{tr}(\boldsymbol{U}_t)}{p}\frac{b_{tij}}{b_{tii}b_{tjj}} = \frac{b_{tij}}{b_{tii}b_{tjj}}. \tag{2.4}$$

To facilitate our analysis, we have assumed $\mathrm{tr}(\boldsymbol{U}_t) = p$ for each session in Equation (2.4) to avoid any identifiability issue, which is formally stated in Assumption 2 of Section 2.3. Note that $\rho_{tij} = -\frac{b_{tij}}{\sqrt{b_{tii}b_{tjj}}} = -r_{tij}\sqrt{b_{tii}b_{tjj}}$. In what follows, we build our test statistics based on some accurate estimation of r_{tij} and the equivalence between $\rho_{tij} = 0$ and $r_{tij} = 0$.

Having discussed the model for each session above, we are in a position to consider all d sessions together to improve the estimation accuracy of each ρ_{tij}. The fact that all sessions tend to share the same support among the column precision matrices $\boldsymbol{R}_t$'s implies the d coefficient vectors $\boldsymbol{\beta}_{ti}$'s share the same support as well. To this end, we natural treat $\beta_{1ij}, ..., \beta_{dij}$ as a group for each pair (i, j), and stack linear models from d sessions together to take the advantage of this group structure. With certain assumption on the joint sparsity structure of d graphs, which is formally stated in Assumption 4 in Section 2.3, the group Lasso ([66]) or other group sparse regression approaches can be adopted to fit our data. We introduce a few more notations before formally stating our procedure. Let $\boldsymbol{\beta}_{ti} = (\beta_{ti1}, \cdots, \beta_{ti(i-1)}, \beta_{ti(i+1)}, \cdots, \beta_{tiq}) \in \mathbb{R}^{q-1}$. The stacked coefficient from d sessions is denoted as $\boldsymbol{\beta}_i^0 = \left(\boldsymbol{\beta}_{1i}{}', \boldsymbol{\beta}_{2i}{}', \cdots, \boldsymbol{\beta}_{di}{}'\right)' \in \mathbb{R}^{(q-1)d}$. By the construction of $\boldsymbol{\beta}_i^0$, we have group sparsity structure in the sense that all but at most s subvectors $\boldsymbol{\beta}_{i(l)}^0$ are none-zero where the lth group subvector of $\boldsymbol{\beta}_i^0$ is defined as $\boldsymbol{\beta}_{i(l)}^0 = (\beta_{1il}, ..., \beta_{dil})' \in \mathbb{R}^d$. From the definition, we observe that $\boldsymbol{\beta}_{i(l)}^0 = \mathbf{0}$ for all $(i, l) \in \mathcal{E}_s^c$, where $\mathcal{E}_s^c$ is the compliment set of $\mathcal{E}_s$.

We are ready to state our regression-based approach with all data $\mathcal{D}$. Let the stacked row samples for each session $\boldsymbol{Z}_t = \left(\boldsymbol{X}_t^{(1)}{}', \boldsymbol{X}_t^{(2)}{}', \cdots, \boldsymbol{X}_t^{(n_t)}{}'\right)' \in \mathbb{R}^{n_t p \times q}$, and its ith column $\boldsymbol{Z}_{t,\cdot,i} = \left(\boldsymbol{X}_{t,\cdot,i}^{(1)}{}', \boldsymbol{X}_{t,\cdot,i}^{(2)}{}', \cdots, \boldsymbol{X}_{t,\cdot,i}^{(n_t)}{}'\right)' \in \mathbb{R}^{n_t p}$. For each session/graph t and each node i, the residuals of the kth sample is denoted as $\boldsymbol{\varepsilon}_{t,\cdot,i}^{(k)} = \boldsymbol{X}_{t,\cdot,i}^{(k)} - \boldsymbol{X}_{t,\cdot,-i}^{(k)}\boldsymbol{\beta}_{ti}$. By combining all d graphs, we estimate the coefficient $\boldsymbol{\beta}_i^0$ using group Lasso,

$$\widehat{\boldsymbol{\beta}}_i^0 = \underset{\boldsymbol{\alpha}_i^0 \in \boldsymbol{R}^{(q-1)d}}{\mathrm{argmin}} \frac{\sum_{t=1}^d \|\boldsymbol{Z}_{t,\cdot,i} - \boldsymbol{Z}_{t,\cdot,-i}\boldsymbol{\alpha}_{ti}\|_2^2}{2n_0 p} + \lambda_i \sum_{l \neq i} \|\boldsymbol{D}_{i(l)}^{1/2}\boldsymbol{\alpha}_{i(l)}^0\|_2, \tag{2.5}$$

where the relationship among $\boldsymbol{\alpha}_{ti}$, $\boldsymbol{\alpha}_i^0$ and $\boldsymbol{\alpha}_{i(l)}^0$ is similar to that among $\boldsymbol{\beta}_{ti}$, $\boldsymbol{\beta}_i^0$ and $\boldsymbol{\beta}_{t(l)}^0$, $n_0 = \min_{1 \leq t \leq d} n_t$ is the smallest sample size among d sessions (see Assumption 1 in Section 2.3 on sample sizes), $\boldsymbol{D}_{ti} \in \mathbb{R}^{(q-1)\times(q-1)}$ is defined as the diagonal matrix of $\frac{\boldsymbol{Z}_{t,\cdot,-i}'\boldsymbol{Z}_{t,\cdot,-i}}{n_t p}$, $\boldsymbol{D}_i \in \mathbb{R}^{(q-1)d\times(q-1)d}$ is the block diagonal matrix with element $(\boldsymbol{D}_{1i}, \boldsymbol{D}_{2i}, \cdots, \boldsymbol{D}_{di})$, and $\boldsymbol{D}_{i(l)} \in \mathbb{R}^{d\times d}$ is the submatrix of $\boldsymbol{D}_i$ corresponding to the lth group. The parameter λ_i can be tuned using cross-validation or other model selection methods. In our theoretical analysis, a data-driven yet conservative choice of each λ_i can be picked. Although group Lasso is not new, our group Lasso regression is applied to, instead of i.i.d. samples, correlated samples since all rows of each matrix $\boldsymbol{X}_t^{(k)}$ are treated as distinct samples. Therefore, in Section 2.3, we provide a self-contained analysis and derive the rates of convergence of estimation as well as prediction, which might be of independent interest for dealing with high-dimensional correlated data.

With estimated regression coefficients, the fitted residuals can be calculated as

$$\widehat{\varepsilon}_{tli}^{(k)} = \boldsymbol{X}_{tli}^{(k)} - \boldsymbol{X}_{tl,-i}^{(k)} \widehat{\boldsymbol{\beta}}_{ti}, \tag{2.6}$$

where $\widehat{\varepsilon}_{tli}^{(k)}$ is the fitted residual for row l of the kth sample for graph t. In order to estimate the population covariance among residuals, the simple empirical covariance using fitted residuals has larger bias compared to the expected root $n_t p$ rate due to the Lasso penalty, as demonstrated in a simpler vector-Gaussian graphical model [38]. To address this bias issue, the covariance of the residuals $\widehat{r}_{tij}$ can be estimated with a bias-correction term as

$$\widehat{r}_{tij} = \begin{cases} -\frac{1}{n_t p} \sum_{k=1}^{n_t} \sum_{l=1}^{p} \left(\widehat{\varepsilon}_{tli}^{(k)} \widehat{\varepsilon}_{tlj}^{(k)} + \widehat{\beta}_{tij} (\widehat{\varepsilon}_{tlj}^{(k)})^2 + \widehat{\beta}_{tji} (\widehat{\varepsilon}_{tli}^{(k)})^2 \right), & \text{if } i \neq j, \\ \frac{1}{n_t p} \sum_{k=1}^{n_t} \sum_{l=1}^{p} \widehat{\varepsilon}_{tli}^{(k)} \widehat{\varepsilon}_{tlj}^{(k)}, & \text{if } i = j. \end{cases} \tag{2.7}$$

In view of the relationship between our goal ρ_{tij} and b_{tij} in Equation (2.4), we can estimate b_{tij} via

$$\widehat{b}_{tij} = \frac{\widehat{r}_{tij}}{\widehat{r}_{tii} \widehat{r}_{tjj}}. \tag{2.8}$$

Finally, the partial correlation ρ_{tij} can be constructed via

$$\widehat{\rho}_{tij} = -\frac{\widehat{b}_{tij}}{(\widehat{b}_{tii} \widehat{b}_{tjj})^{1/2}}. \tag{2.9}$$

Our estimator for partial correlation resembles the form proposed by [16] for a single matrix-variate Gaussian graphical model. However, in contrast with [16], we borrow the information from multiple sessions and apply a group Lasso in Equation (2.5) to obtain a faster rate of convergence for estimating the coefficient of each session/graph, which is summarized in Theorem 1 and Remark 3. More importantly, as we introduce the final form of test statistic in Equation (2.10), the testing accuracy for our goals can be further improved with a factor of root d, thanks to the similar graph structures among multiple sessions.

2.2.2 Simultaneous Test by Parametric Bootstrap

Single Edge Test

The way of stacking correlated rows of each matrix-variate sample and combining multiple graphs in Equation (2.5) not only enables us to estimate d graphs efficiently through a group Lasso regression, but also allows us to construct new tests that are more powerful than the single graph scenario by borrowing information from different graphs. We first focus on the single edge test $H_{0,ij} : \boldsymbol{\rho}_{ij}^0 = \mathbf{0}$ in Equation (2.1). Due to the group sparsity structure, it is natural to construct a test statistic using certain function of all estimators $\widehat{\rho}_{tij}$ for $1 \leq t \leq d$. In our application, we assume that the sign does not change across all sessions, which corresponds to a sign vector of length d where all elements equal to the same value one. More generally, an additional sign information on elements of the alternative $\boldsymbol{\rho}_{ij}^0$ may be available. With this additional knowledge, we present a test statistic based on a linear combination of those $\widehat{\rho}_{tij}$ for $1 \leq t \leq d$, which is

closely related to its ℓ_1 norm. More specifically, with a sign vector $\boldsymbol{\xi}_{ij} = (\xi_{1ij}, \ldots, \xi_{dij})' \in \mathbb{R}^d$, we propose the following test statistic for our hypothesis testing problem in Equation (2.1),

$$P_{n,d,(i,j)}(\boldsymbol{\xi}) = -\frac{1}{\sqrt{d}} \sum_{t=1}^{d} \xi_{tij} \sqrt{n_t p \widehat{b}_{tii} \widehat{b}_{tjj}} \widehat{r}_{tij} = \frac{1}{\sqrt{d}} \sum_{t=1}^{d} \xi_{tij} \sqrt{n_t p} \widehat{\rho}_{tij}. \tag{2.10}$$

Intuitively, this test statistic should be close to its population counterpart $P^*_{n,d,(i,j)}(\boldsymbol{\xi}) = \frac{1}{\sqrt{d}} \sum_{t=1}^{d} \xi_{tij} \sqrt{n_t p} \rho_{tij}$, which is equal to zero under null. On the other hand, when alternative is true, with the sign vector $\xi_{tij} = \text{sign}(\rho_{tij})$ for $1 \leq t \leq d$, $P^*_{n,d,(i,j)}(\boldsymbol{\xi})$ should be significantly larger than zero as it represents the aggregated signal across d graphs in terms of the ℓ_1 norm of $\boldsymbol{\rho}^0_{ij}$. Indeed, as we formally show in Proposition 1 of Section 2.3, the difference $\widehat{\rho}_{tij} - \rho_{tij}$ under certain sample size requirement can be well approximated by a leading term

$$\theta_{tij} = \frac{\widetilde{\delta}_{tij}}{\sqrt{r_{tii} r_{tjj}}} - \frac{r_{tij} \widetilde{\delta}_{tjj}}{2 r_{tjj} \sqrt{r_{tii} r_{tjj}}} - \frac{r_{tij} \widetilde{\delta}_{tii}}{2 r_{tii} \sqrt{r_{tii} r_{tjj}}}, \tag{2.11}$$

where $\widetilde{\delta}_{tij} = \widetilde{r}_{tij} - r_{tij}$ is the error of the oracle estimator of r_{tij}, $\widetilde{r}_{tij} = \frac{1}{n_t p} \sum_{k=1}^{n_t} \sum_{l=1}^{p} \varepsilon^{(k)}_{tli} \varepsilon^{(k)}_{tlj}$. Although residuals $\varepsilon^{(k)}_{tli}$'s from individual rows $1 \leq l \leq p$ are correlated, by certain version of central limit theorem it is well expected that the following difference $\Delta P_{i,j}$ is asymptotically close to a normal distribution with a finite variance,

$$\Delta P_{i,j} = P_{n,d,(i,j)}(\boldsymbol{\xi}) - P^*_{n,d,(i,j)}(\boldsymbol{\xi}) = \frac{1}{\sqrt{d}} \sum_{t=1}^{d} \xi_{tij} \sqrt{n_t p} (\widehat{\rho}_{tij} - \rho_{tij}). \tag{2.12}$$

Once the asymptotic variance can be consistently estimated, it is straightforward to construct the confidence interval for $P^*_{n,d,(i,j)}(\boldsymbol{\xi})$ and the p-value for our single edge test in Equation (2.1). We do not pursue this direction immediately here. Instead, as we move to the more challenging multiple edge test scenario in Equation (2.2), one can easily obtain the testing procedure for the single edge test as it is just a special case of multiple edge test.

Simultaneous Test

We consider the multiple edge test for a general pre-defined set $\mathcal{S}$ with the null hypothesis stated in Equation (2.2),

$$H_{0,\mathcal{S}} : \boldsymbol{\rho}^0_{ij} = \mathbf{0}, \ \forall (i,j) \in \mathcal{S}.$$

In high-dimensional setting, the cardinality of $\mathcal{S}$ can be as large as $q(q-1)/2$. In practice, $\mathcal{S}$ can be a collection of edges between different brain areas. Therefore, even if we have asymptotic normality for each $\Delta P_{i,j}$, $(i,j) \in \mathcal{S}$, the multivariate central limit theorem, which is valid for fixed $|\mathcal{S}|$, cannot be applied to test $H_{0,\mathcal{S}}$ directly with growing $|\mathcal{S}|$. Simultaneous testing for $H_{0,\mathcal{S}}$ has been considered in other context earlier via Bonferroni correction, which leads to conservative procedures. In this paper, due to the sparsity structure on the entire graphs, we consider a test statistic using the maximum among all $\Delta P_{i,j}$'s in magnitude. More specifically, for a given edge set $\mathcal{S} \subset \{(i,j) : 1 \leq i \neq j \leq q\}$ with size $e = |\mathcal{S}|$, denote the e-dimensional vector by $\Delta \boldsymbol{P}_{\mathcal{S}}$

whose elements correspond to the test statistics for the subset of edges of interest. Now we can define a bijective mapping $\boldsymbol{\chi}(\cdot) = \{\chi_1(\cdot), \chi_2(\cdot)\}$ from $\{1, ..., e\}$ to $\mathcal{S}$ such that we can write $\Delta \boldsymbol{P}_{\mathcal{S}} = \{\Delta P_{\chi_1(1),\chi_2(1)}, ..., \Delta P_{\chi_1(e),\chi_2(e)}\}$. Then our test statistic is defined as

$$\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty = \max_{(i,j)\in\mathcal{S}} |\Delta P_{i,j}| = \max_{(i,j)\in\mathcal{S}} |\frac{1}{\sqrt{d}} \sum_{t=1}^{d} \xi_{tij} \sqrt{n_t p}(\widehat{\rho}_{tij} - \rho_{tij})|, \tag{2.13}$$

where the sign vector is edge-specific, i.e., $\boldsymbol{\xi}_{ij} = (\xi_{1ij}, \ldots, \xi_{dij})' \in \mathbb{R}^d$ for each pair $(i, j) \in \mathcal{S}$.

The key idea is that although one do not have asymptotic normal for the entire vector $\Delta \boldsymbol{P}_{\mathcal{S}}$, following the idea of [17], it can be shown that the limiting behavior of $\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty$ can be approximated by that of the supnorm of a certain multivariate normal vector. We will elucidate this idea later. Based on our test statistic, a $1 - \alpha$ confidence region for partial correlation vector can be constructed as

$$C_S(1-\alpha) = \{\boldsymbol{c} \in \mathbb{R}^{d|\mathcal{S}|} : \max_{(i,j)\in\mathcal{S}} |\frac{\sum_{t=1}^{d} \xi_{tij}\sqrt{n_t p}(\widehat{\rho}_{tij} - c_{tij})}{\sqrt{d}}| \leq \widehat{q}_{\boldsymbol{\zeta}}(\alpha)\}, \tag{2.14}$$

with some well estimated critical value $\widehat{q}_{\boldsymbol{\zeta}}(\alpha)$. Given the test level α for $H_{0,\mathcal{S}}$, we reject the null hypothesis if $\boldsymbol{0} \notin C_S(1-\alpha)$. In other words, we reject null hypothesis if $\Psi_\alpha = 1$ where

$$\Psi_\alpha = \mathbb{1}(\boldsymbol{0} \notin C_S(1-\alpha)). \tag{2.15}$$

It remains to provide a critical value $\widehat{q}_{\boldsymbol{\zeta}}(\alpha)$ for our testing procedure. As we discussed for single edge test, the leading term of the difference $\widehat{\rho}_{tij} - \rho_{tij}$ under certain sample size requirement is denoted by θ_{tij} in Equation (2.11). Similar to the way we defined $\boldsymbol{\chi}(\cdot)$, we represent $\boldsymbol{\Theta}_{t\mathcal{S}} = \{\theta_{t,\chi_1(1),\chi_2(1)}, ..., \theta_{t,\chi_1(e),\chi_2(e)}\}$ and $\boldsymbol{\xi}_{t\mathcal{S}} = \{\xi_{t,\chi_1(1),\chi_2(1)}, ..., \xi_{t,\chi_1(e),\chi_2(e)}\}$. At a high level, the distribution of $\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty$ should be close to that of $\|\sum_{t=1}^{d} \frac{\sqrt{n_t p}}{\sqrt{d}} \boldsymbol{\xi}_{t\mathcal{S}} \circ \boldsymbol{\Theta}_{t\mathcal{S}}\|_\infty$, which is further close to the supnorm of a multivariate normal with the same covariance matrix of $\sum_{t=1}^{d} \frac{\sqrt{n_t p}}{\sqrt{d}} \boldsymbol{\xi}_{t\mathcal{S}} \circ \boldsymbol{\Theta}_{t\mathcal{S}}$, where $\circ$ is the Hadamard product. This is formally stated and shown in Proposition 2. In order to use this approximation to obtain critical value $\widehat{q}_{\boldsymbol{\zeta}}(\alpha)$, we need to know the covariance matrix $\boldsymbol{W}^P$, which is defined as

$$\boldsymbol{W}^P = \mathbb{E}\{(\frac{1}{\sqrt{d}} \sum_{t=1}^{d} \sqrt{n_t p} \boldsymbol{\xi}_{t\mathcal{S}} \circ \boldsymbol{\Theta}_{t\mathcal{S}})(\frac{1}{\sqrt{d}} \sum_{t=1}^{d} \sqrt{n_t p} \boldsymbol{\xi}_{t\mathcal{S}} \circ \boldsymbol{\Theta}_{t\mathcal{S}})'\}. \tag{2.16}$$

An analytical form for $\boldsymbol{W}^P = (w_{ij}^P)_{e\times e}$ is provided in Lemma 11 as

$$\begin{aligned}
w_{ij}^P = \frac{1}{d} \sum_{t=1}^{d} \big(& \frac{\|\boldsymbol{U}_t\|_F^2}{p} (\rho_{t,\chi_1(i),\chi_1(j)} \rho_{t,\chi_2(i),\chi_2(j)} + \rho_{t,\chi_1(i),\chi_2(j)} \rho_{t,\chi_1(j),\chi_2(i)} \\
& + \frac{1}{2} \rho_{t,\chi_1(i),\chi_2(i)} \rho_{t,\chi_1(j),\chi_2(j)} \rho_{t,\chi_1(i),\chi_1(j)}^2 + \frac{1}{2} \rho_{t,\chi_1(i),\chi_2(i)} \rho_{t,\chi_1(j),\chi_2(j)} \rho_{t,\chi_2(i),\chi_2(j)}^2 \\
& + \frac{1}{2} \rho_{t,\chi_1(i),\chi_2(i)} \rho_{t,\chi_1(j),\chi_2(j)} \rho_{t,\chi_1(i),\chi_2(j)}^2 + \frac{1}{2} \rho_{t,\chi_1(i),\chi_2(i)} \rho_{t,\chi_1(j),\chi_2(j)} \rho_{t,\chi_1(j),\chi_2(i)}^2 \\
& - \rho_{t,\chi_1(i),\chi_1(j)} \rho_{t,\chi_1(j),\chi_2(j)} \rho_{t,\chi_1(j),\chi_2(i)} - \rho_{t,\chi_1(i),\chi_1(j)} \rho_{t,\chi_1(i),\chi_2(i)} \rho_{t,\chi_1(i),\chi_2(j)} \\
& - \rho_{t,\chi_2(i),\chi_2(j)} \rho_{t,\chi_1(j),\chi_2(j)} \rho_{t,\chi_1(i),\chi_2(j)} - \rho_{t,\chi_2(i),\chi_2(j)} \rho_{t,\chi_1(j),\chi_2(i)} \rho_{t,\chi_1(i),\chi_2(i)}) \big).
\end{aligned} \tag{2.17}$$

Due to the temporal structure, $\boldsymbol{W}^P$ depends on the temporal convariance matrices $\boldsymbol{U}_t$'s via their Frobenius norms. In next subsection, we develop a procedure based on modified Cholesky decomposition of $\boldsymbol{A}_t$ to obtain a consistent estimator $\widehat{\boldsymbol{U}}_t^*$ in Equation (2.24) under the Frobenius norm. With $\widehat{\boldsymbol{U}}_t^*$'s and $\widehat{\rho}_{tij}$'s obtained in Equation (2.9), we can construct a plug-in estimator $\widehat{\boldsymbol{W}}^P$ using Equation (2.17).

Now let $\widehat{\boldsymbol{\zeta}}_1, ..., \widehat{\boldsymbol{\zeta}}_B$ be i.i.d random vector with each sample $\widehat{\boldsymbol{\zeta}}_i \sim N(\boldsymbol{0}, \widehat{\boldsymbol{W}}^P)$, we can approximate the distribution of $\|\widehat{\boldsymbol{\zeta}}\|_\infty$ using the empirical distribution of B bootstrap samples, $\widehat{F}_{\boldsymbol{\zeta},B}(x) = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}\{\|\widehat{\boldsymbol{\zeta}}_i\|_\infty \leq x\}$. Finally, the critical value $\widehat{q}_S(\alpha)$ used in our testing procedure Equations (2.14)-(2.15) can be estimated by the quantile function of $\widehat{F}_{\boldsymbol{\zeta},B}(x)$ as,

$$\widehat{q}_S(\alpha) = \inf\{x \in \mathbb{R} : \widehat{F}_{\boldsymbol{\zeta},B}(x) \geq 1 - \alpha\}.$$

The validity of our Gaussian approximation is formally provided in Theorem 2 in terms of the Kolmogorov distance between the distributions of $\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty$ and $\|\widehat{\boldsymbol{\zeta}}\|_\infty$. In addition, we also provide a complementary power analysis in Theorem 3 to further demonstrate the advantage of our procedure with multiple graphs.

Remark 1. *With matrix-variate data, we have correlated samples in the leading term θ_{tij} of our test statistic, where the correlation is characterized by the temporal covariance matrix $\boldsymbol{U}_t$. Thanks to the Kronecker product structure, we can consistently estimate $\boldsymbol{W}^P$ with a plug-in estimator. It is worthwhile to mention that under other dependence structures among samples, one may construct different procedures for a vector-variate Gaussian graph. For instance, [15] implemented a kernel estimator for estimation of $\boldsymbol{W}^P$ in multi-variate time series data.*

We point out that the proposed multiple edge test can be extended straightforwardly to a general simultaneous testing $H_{0,\mathcal{S}} : \boldsymbol{\rho}_{ij}^0 = \boldsymbol{c}_{ij}^0, \forall (i,j) \in \mathcal{S}$ where $\boldsymbol{c}_{ij}^0 = (c_{1ij}, \ldots, c_{dij})'$ is a pre-specified vector for each edge $(i,j) \in \mathcal{S}$. The interpretation of the sign vector should be the sign of the difference between $\boldsymbol{c}_{ij}^0$ and alternative $\boldsymbol{\rho}_{ij}^0$, i.e., $\xi_{tij} = \text{sign}(\rho_{tij} - c_{tij})$ when alternative is true. Denote $\boldsymbol{c} = \{c_{tij} : (i,j) \in \mathcal{S}, 1 \leq t \leq d\}$. One only need to replace the construction of confidence region in Equation (2.15) by $\Psi_\alpha = \mathbb{1}(\boldsymbol{c} \notin C_S(1-\alpha))$. We present our theoretical justifications in Section 2.3 based on this more general procedure.

Finally, we extend our current simultaneous testing to a c-level test as earlier discussed in Equation (2.3), we are interested in testing

$$H'_{0,\mathcal{S}} : \frac{1}{d}\|\boldsymbol{\rho}_{ij}^0\|_1 \leq c \text{ for all } (i,j) \in \mathcal{S}.$$

Under our assumption that all $\rho_{1ij}, \ldots, \rho_{dij}$ share the same sign, the null can be equivalently written as $H'_{0,\mathcal{S}} : \frac{1}{d}|\sum_{t=1}^{d} \rho_{tij}| \leq c$ and $\text{sign}(\rho_{1ij}) = \ldots = \text{sign}(\rho_{dij})$ for all $(i,j) \in \mathcal{S}$. One can still obtain a confidence region from Equation (2.14) with $\xi_{tij} = 1$ for all $1 \leq t \leq d$ and $(i,j) \in \mathcal{S}$. However, since the null hypothesis is composite now, we should modify the rejection region accordingly. To this end, define $\boldsymbol{C}_{\text{c-level}} = \{\boldsymbol{c} \in \mathbb{R}^{d|\mathcal{S}|} : \frac{1}{d}|\sum_{t=1}^{d} c_{tij}| \leq c \text{ and } \text{sign}(c_{1ij}) = \ldots = \text{sign}(c_{dij}) \text{ for all } (i,j) \in \mathcal{S}\}$

Then we reject null hypothesis if $\Psi'_\alpha = 1$ where

$$\Psi'_\alpha = \mathbb{1}(\boldsymbol{C}_{\text{c-level}} \cap C'_S(1-\alpha)) \neq \emptyset). \tag{2.18}$$

2.2.3 Estimation of Temporal Covariance Matrix

Our testing framework developed in Section 2.2.2 requires the knowledge of the temporal covariance $\boldsymbol{U}_t$ in terms of its Frobenius norm $\|\boldsymbol{U}_t\|_F$. In this subsection, we propose an estimation procedure based on modified Cholesky decomposition. Motivated by our data, unlike the way we estimate the spatial precision matrices borrowing the information across all sessions in Section 2.2.1, here we will estimate $\boldsymbol{U}_t$ for each graph/session individually only based on certain bandable structure imposed on the Cholesky factor of each $\boldsymbol{A}_t = \boldsymbol{U}_t^{-1}$. Indeed, despite that the experimental stages are properly aligned, the temporal alignment of neural response is still an open research question due to the existence of response latencies of neurons to stimulus [58]. Especially, the neural response time for saccade tasks might depend on a variety of factors in previous studies [2, 21, 47], such as saccade amplitude, direction and change in luminance. As the neural reaction time to stimulus may be impacted by even more complicated factors across experimental sessions, we do not assume $\boldsymbol{U}_t$'s are identical or very close to each other, but rather simply assume all temporal precision matrices $\boldsymbol{A}_t$'s demonstrate certain structure, which would be introduced below.

Our procedure is motivated by the temporal dependence structure of neural time series and based on modified Cholesky decomposition [5, 40]. Neural time series such as the Local Field Potential have been widely modeled as an autoregression problem with a limited order [8, 9, 55]. Assume that $\boldsymbol{Y}_t = (Y_{t1}, ..., Y_{tp})^{'}$ is a centered p-variate random vector with temporal covariance $\boldsymbol{U}_t$, Y_{ti} can be predicted using the measurements from a few previous timeticks. Now assume that $\boldsymbol{c}_{ti} = (c_{t1}, ..., c_{t(i-1)})'$ is the coefficient for the following population regression,

$$\widehat{Y}_{ti} = \sum_{j=1}^{i-1} c_{tij} Y_{tj} = \boldsymbol{Y}_{t,1:i-1}^{'} \boldsymbol{c}_{ti}, \tag{2.19}$$

where $\boldsymbol{Y}_{t,1:i-1} = (Y_{t,1}, Y_{t,2}, \cdots, Y_{t,i-1})^{'}$. Let $\boldsymbol{\epsilon}_t = \boldsymbol{Y}_t - \widehat{\boldsymbol{Y}}_t$ denote the residual, $d_{ti} = \text{Var}(\epsilon_{ti})$ denote variance of residual, $\boldsymbol{D}_t = \text{diag}(\boldsymbol{d}_t)$ denote the diagonal matrix of residual, $\boldsymbol{C}_t$ denote the lower triangular matrix with zeros on the diagonal and the coefficients $\boldsymbol{c}_{ti}$ arranged in row i. Then the Cholesky decomposition of $\boldsymbol{A}_t = \boldsymbol{U}_t^{-1}$ can be written as

$$\boldsymbol{A}_t = (\boldsymbol{I} - \boldsymbol{C}_t)^{'} \boldsymbol{D}_t^{-1} (\boldsymbol{I} - \boldsymbol{C}_t). \tag{2.20}$$

In high-dimensional neural time series setting, the dependence of Y_{ti} on its history $Y_{tj} (j < i)$ grows weaker and exhibits natural decay as $i - j$ becomes larger, which indicates that c_{tij} becomes smaller accordingly. In this paper, we consider a parameter space for each $\boldsymbol{A}_t$ as formally stated in Assumption 6 later,

$$\mathcal{Q}_{\alpha_t}(M) = \{\boldsymbol{A}_t : |c_{tij}| < M(i-j)^{-\alpha_t - 1}, 1 \leq j \leq i-1\},$$

where the value α_t specifying the rate of decay is assumed to be known. In fact, the popular AR(k) model corresponds to the k-banded Cholesky factor decomposition where $c_{tij} = 0$ if $i - j > k$.

The above model focuses on the standard vector-variate distributions. For our matrix-variate Gaussian graphical model, we can apply each column of $\boldsymbol{X}_t \in \mathbb{R}^{p \times q}$ to obtain the Cholesky decomposition for the corresponding node. Similar to the observation we made on node-wise regression for each row of $\boldsymbol{X}_t$ in Section 2.2.1, due to the Kronecker product structure, the

population Cholesky factor $\boldsymbol{C}_t$ remains the same while the diagonal matrices $\boldsymbol{D}_t$'s are proportional to each other for regression models across all q columns of $\boldsymbol{X}_t$. This motivates us to treat all columns of $\boldsymbol{X}_t$ as individual samples. By doing this, the actual number of samples can be significantly increased, thanks to the matrix-variate data structure. Of note, those samples obtained from individual columns of $\boldsymbol{X}_t$ are no longer independent and their covariance is characterized by the spatial covariance matrix $\boldsymbol{V}_t$. To formally introduce our estimation procedure, for each session t, we denote $\boldsymbol{Z}_{t,i,\cdot} = \left(\boldsymbol{X}_{t,i,\cdot}^{(1)}, \boldsymbol{X}_{t,i,\cdot}^{(2)}, \cdots, \boldsymbol{X}_{t,i,\cdot}^{(n_t)}\right)' \in \mathbb{R}^{n_t q}$, and $\boldsymbol{Z}_{t,i-h:i-1,\cdot} = (\boldsymbol{Z}_{t,i-h,\cdot}, \boldsymbol{Z}_{t,i-h+1,\cdot}, \cdots, \boldsymbol{Z}_{t,i-1,\cdot}) \in \mathbb{R}^{n_t q \times p}$. We follow the procedure proposed in [40] for i.i.d samples and apply it to our correlated samples. The first step is to find the empirical regression coefficients $\boldsymbol{c}_{ti}$ by least squares for the ith variable only against its previous h_t number of variables instead of all variables shown in the true regression model in Equation (2.19). The idea is that by ignoring many weak signals far away from variables i, the estimation of $\boldsymbol{c}_{ti}$ can be achieved optimally with the bias-variance trade-off. Specifically, define bandwidth $h_t = \lceil (n_t q)^{\frac{1}{2\alpha_t+2}} \rceil$, we have

$$\widehat{\boldsymbol{c}}_{ti}^{h} = (\widehat{c}_{ti(i-h)}, ..., \widehat{c}_{ti(i-1)})' = (\boldsymbol{Z}_{t,i-h:i-1,\cdot}{}' \boldsymbol{Z}_{t,i-h:i-1})^{-1} \boldsymbol{Z}_{t,i-h:i-1,\cdot}{}' \boldsymbol{Z}_{t,i,\cdot}. \tag{2.21}$$

Let $\widehat{\boldsymbol{c}}_{ti}$ be a vector of length $i-1$ with the first $i-1-h$ elements padded as zero, and other entries filled with $\widehat{\boldsymbol{c}}_{ti}^{h}$. Then $\boldsymbol{C}_t$ can be estimated by the lower triangular matrix $\widehat{\boldsymbol{C}}_t$ with zeros on the diagonal and the coefficients $\widehat{\boldsymbol{c}}_{ti}$ arranged in row i. Having obtained the empirical regression coefficients $\widehat{\boldsymbol{c}}_{ti}^{h}$, the second step is to estimate the noise variance given by

$$\widehat{d}_{ti} = \frac{1}{n_t q} \sum_{k=1}^{n_t} \sum_{m=1}^{q} \left(X_{tim}^{(k)} - (\boldsymbol{X}_{t,i-h:i-1,m}^{(k)})' \widehat{\boldsymbol{c}}_{ti}^{h} \right)^2. \tag{2.22}$$

Let $\widehat{\boldsymbol{D}}_t = \mathrm{diag}(\widehat{\boldsymbol{d}}_t) \in \mathbb{R}^{p \times p}$, where $\widehat{\boldsymbol{d}}_t = (\widehat{d}_{t1}, ..., \widehat{d}_{tp})'$. The final estimator of $\boldsymbol{U}_t$ is defined as

$$\widehat{\boldsymbol{U}}_t = \widehat{\boldsymbol{A}}_t^{-1} = \left((\boldsymbol{I} - \widehat{\boldsymbol{C}}_t)' \widehat{\boldsymbol{D}}_t^{-1} (\boldsymbol{I} - \widehat{\boldsymbol{C}}_t) \right)^{-1}. \tag{2.23}$$

As shown in Theorem 4 of Section 2.3, our estimator $\widehat{\boldsymbol{U}}_t$ actually aims to the normalized temporal covariance matrix $\frac{\mathrm{Tr}(\boldsymbol{V}_t)}{q} \boldsymbol{U}_t$ rather than $\boldsymbol{U}_t$. In view of Assumption 2 to avoid the identifiability issue, finally we define $\widehat{T}_v = \frac{Tr(\widehat{\boldsymbol{U}}_t)}{p}$, and the normalized estimator for temporal covariance matrix is given by

$$\widehat{\boldsymbol{U}}_t^{*} = \frac{1}{\widehat{T}_v} \widehat{\boldsymbol{U}}_t. \tag{2.24}$$

The theoretical properties of our estimator $\widehat{\boldsymbol{U}}_t^{*}$ under correlated samples for each session are deferred to Theorem 4 and Corollary 2.

2.3 Theoretical Properties

Before formally presenting our procedure and main results, we make the following assumptions for our model.

Algorithm 1 Simultaneous Testing for Multiple MGGMs

Input: Multi-session data $\mathcal{D}$, edge set $\mathcal{S}$, test level α.
Output: Test result Ψ_α.
Spatial precision matrix estimation:
for $i = 1 : q$ **do**
    Estimate the regression coefficient $\widehat{\boldsymbol{\beta}}_i^0$ and residual using Equation (2.5).
    Estimate the regression residual $\widehat{\varepsilon}_{tli}^{(k)}$ using Equation (2.6).
end for
for $t = 1 : d$, $i = 1 : q$, $j = 1 : q$ **do**
    Estimate the de-biased residual variance $\widehat{r}_{tij}$ using Equation (2.7).
    Estimate the spatial precision $\widehat{b}_{tij}$ and partial correlation $\widehat{\rho}_{tij}$ using
    Equation (2.8) and (2.9).
end for
Temporal precision matrix estimation:
for $t = 1 : d$ **do**
    Estimate temporal regression coefficient and residual in Equations (2.21), (2.22).
end for
Hypothesis testing based on bootstrap:
Calculate the plug-into estimator $\widehat{\boldsymbol{W}}^P$ for bootstrap covariance in Equation (2.17).
Sample $\left\{\widehat{\boldsymbol{\zeta}}_i\right\}_{i=1}^{B} \sim N(\mathbf{0}, \widehat{\boldsymbol{W}}^P)$ and calculate the confidence region by Equation (2.14).
Test result Ψ_α is given by Equation (2.15).
return test results Ψ_α.

Assumption 1. *It holds that* $n_1 \asymp ... \asymp n_t$ *with* $\asymp$ *meaning asymptotically the same order, and that* $\max_{1\le t\le d} \frac{n_t}{n_0} \le M_0$, $n_0 = \min_{1\le t\le d} n_t$, *where* M_0 *is a positive constant.*

Assumption 2. *We assume that* $\frac{\mathrm{tr}(\boldsymbol{U}_t)}{p} = 1$ *for all* $1 \le t \le d$.

Assumption 3. *Define* $\{\lambda_{ti}\}_{i=1}^p$ *as the eigenvalues of* $\boldsymbol{U}_t$ *with* $\lambda_{t1} \le \lambda_{t2} \cdots \le \lambda_{tp}$, *and* $\{\lambda'_{ti}\}_{i=1}^q$ *as the eigenvalues of* $\boldsymbol{V}_t$ *with* $\lambda'_{t1} \le \lambda'_{t2} \cdots \le \lambda'_{tq}$; *there exists a constant* c_e, *so that*

$$c_e^{-1} \le \lambda_{t1} \le \lambda_{tp} \le c_e \text{ and } c_e^{-1} \le \lambda'_{t1} \le \lambda'_{tq} \le c_e \text{ for all } 1 \le t \le d.$$

Assumption 4. *The* l*-th group structure of* $\{\boldsymbol{B}_t\}_{t=1}^d$ *is defined as* $\boldsymbol{b}_{i(l)} = (b_{1il}, ..., b_{dil}) \in \mathbb{R}^d$; *the group sparsity* s *is defined as the maximum node degree, i.e.,* $s = \max_i \sum_{l\ne i}\{b_{i(l)} \ne \boldsymbol{0}\}$ *with* s *satisfying*

$$\frac{\max\left(s^2 \log q(\log q + d)^2, \log^7 q\right)}{n_0 p d} = o(1).$$

Assumption 5. *There are some constants* $\delta_1 = o(1)$, *with* $\log(d/\delta_1) = O(s\frac{d+\log q}{d})$.

Assumption 6. *We assume the temporal precision matrix* $\boldsymbol{A}_t$ *for each* $1 \le t \le d$ *belongs to the following parameter space*

$$\mathcal{Q}_{\alpha_t}(M) = \{\boldsymbol{A}_t : |c_{tij}| < M(i-j)^{-\alpha_t - 1}, 1 \le j \le i-1\},$$

where $\boldsymbol{C}_t$ *is defined in Equation* (2.20) *through the Cholesky decomposition. We further assume that* $\frac{\log \max(p, n_t q)}{(n_t q)^{\frac{\alpha_t}{\alpha_t+1}}} = o(1)$.

Assumption 1 suggests that the sample size from each graph is balanced and we use n_0 to represent the common level; Assumption 2 is simply for identifiability; Assumption 3 is a standard eigenvalue assumption in covariance estimation [13]; Assumption 4 indicates that the column precision matrix is sparse, and limits the spatial dimension of the matrix variable given the number of samples, the temporal dimension and the number of graphs; Assumption 5 limits the number of graphs with respect to the spatial dimension and sparsity; the first part of Assumption 6 is a fair assumption for neural time series, since neural data, especially LFPs, are usually modeled as an auto-regressive process with limited order, which is also widely considered in literature [5, 40]; the second part of Assumption 6 is similar to Assumption 4 and limits the temporal dimension.

We first provide a theoretical justification for our group Lasso procedure proposed in Section 2.2.1. Although with correlated rows, our results below demonstrate that the optimal rates of convergence for estimation and prediction can be still obtained compared to the case with i.i.d. samples. Define $\lambda_* = \sqrt{\frac{d+\log q}{n_0 p}}$; we have the following theorem for our regression coefficients in Equation (2.5),

Theorem 1. *Let* $\lambda_i = c\frac{\xi+1}{\xi-1}\sqrt{\frac{d+\log q}{n_0 p}} = c\frac{\xi+1}{\xi-1}\lambda_*$ *as in Equation* (2.5), *with positive constants* $c > 0$ *and* $\xi > 1$. *With probability at least* $1 - q^{-\delta}$, *for all* $1 \le i \le q$, *there is a constant* C *which depends on* δ, c_e *and* M_0 *only, we have that*

$$\sum_{1\le l\le q, l\ne i} \frac{1}{\sqrt{d}} \|\widehat{\boldsymbol{\beta}}^0_{i(l)} - \boldsymbol{\beta}^0_{i(l)}\|_2 \le Cs[\frac{1 + (\log q)/d}{n_0 p}]^{1/2}, \tag{2.25}$$

$$\frac{1}{\sqrt{d}} \|\widehat{\boldsymbol{\beta}}^0_{i} - \boldsymbol{\beta}^0_{i}\|_2 \le C[s\frac{1 + (\log q)/d}{n_0 p}]^{1/2}, \tag{2.26}$$

$$\frac{1}{n_0 pd}\sum_{t=1}^{d}\|\boldsymbol{Z}_{t,\cdot,-i}(\widehat{\boldsymbol{\beta}}_{ti}-\boldsymbol{\beta}_{ti})\|_2^2 \le Cs\frac{1+(\log q)/d}{n_0 p}. \tag{2.27}$$

Remark 2. *In Lemma 3, we provide a data-driven yet conservative method to pick the proper λ_i, so that the major conclusions in this paper still hold under this tuning parameter. In practice, we can also implement the cross-validation method to pick λ_i in group Lasso regression. Moreover, we found that the performance of our method is much better than baseline methods regardless of tuning parameters, especially under the condition which favors our method, i.e., under high temporal and spatial dimensions, as demonstrate in the simulation in Section 2.4.*

We now turn to theoretical results from Section 2.2.2. Based on Theorem 1, the following proposition provides an upper bound for the remainder term of each $\widehat{\rho}_{tij}-\rho_{tij}-\theta_{tij}$, where the estimator $\widehat{\rho}_{tij}$ is proposed in Equation (2.9) and $\theta_{tij}=\frac{\widetilde{\delta}_{tij}}{\sqrt{r_{tii}r_{tjj}}}-\frac{r_{tij}\widetilde{\delta}_{tjj}}{2r_{tjj}\sqrt{r_{tii}r_{tjj}}}-\frac{r_{tij}\widetilde{\delta}_{tii}}{2r_{tii}\sqrt{r_{tii}r_{tjj}}}$.

Proposition 1. *we have for any $\delta>0$, and δ_1 as in Assumption 5, there exists a constant C which depends on δ, c_e and M_0 only, such that*

$$\mathbb{P}\left(\sum_{t=1}^{d}|\widehat{\rho}_{tij}-\rho_{tij}-\theta_{tij}|\ge Cs\lambda_*^2\right)\le q^{-\delta}+\delta_1.$$

It then immediately follows that $\Delta\boldsymbol{P}_{\mathcal{S}}=\boldsymbol{P}_{n,d,\mathcal{S}}(\boldsymbol{\xi})-\boldsymbol{P}^*_{n,d,\mathcal{S}}(\boldsymbol{\xi})=\frac{1}{\sqrt{d}}\sum_{t=1}^{d}\sqrt{n_t p}(\boldsymbol{\xi}_{t\mathcal{S}}\circ\boldsymbol{\Theta}_{t\mathcal{S}}+\boldsymbol{\xi}_{t\mathcal{S}}\circ\boldsymbol{O}_{t\mathcal{S}})$, with $\|\sum_{t=1}^{d}\frac{\sqrt{n_t p}}{\sqrt{d}}\boldsymbol{\xi}_{t\mathcal{S}}\circ\boldsymbol{O}_{t\mathcal{S}}\|_\infty$ being a smaller term as shown in Proposition 1.

Remark 3. *For single edge test, we need to make $\sum_{t=1}^{d}\frac{\sqrt{n_t p}}{\sqrt{d}}\boldsymbol{\xi}_{t\mathcal{S}}\circ\boldsymbol{O}_{t\mathcal{S}}$ an $o_p(1)$ term, which imposes the condition that $\frac{s^2(d+\log q)^2}{d}=o(n_0 p)$; for multiple edge test (simultaneous test), in order to make $\|\sum_{t=1}^{d}\frac{\sqrt{n_t p}}{\sqrt{d}}\boldsymbol{\xi}_{t\mathcal{S}}\circ\boldsymbol{O}_{t\mathcal{S}}\|_\infty$ an $o_p(1)$ term, the sample size requirement is $\frac{s^2\log q(d+\log q)^2}{d}=o(n_0 p)$, as stated in Assumption 4. In comparison, one can also naively apply the estimation procedure for each graph separately as in [16], and compute a similar test statistic to perform single edge and multiple edge test following our procedure. However, such a naive method will require a much stronger sample size assumption, which is $s^2 d(\log q)^2=o(n_0 p)$ in single edge test and $s^2 d(\log q)^3=o(n_0 p)$ in simultaneous test.*

Built upon the idea from [17], we establish the Gaussian approximation result for our test statistic $\|\Delta\boldsymbol{P}_{\mathcal{S}}\|_\infty$. Recall the covariance from the leading term $\boldsymbol{W}^P$ is defined in Equation (2.16).

Proposition 2. *Let $\boldsymbol{\zeta}\sim N(\boldsymbol{0},\boldsymbol{W}^P)$ where $\boldsymbol{W}^P$ is defined in Equation* (2.16)*, it holds that*

$$\sup_{x>0}|\mathbb{P}(\|\Delta\boldsymbol{P}_{\mathcal{S}}\|_\infty>x)-\mathbb{P}(\|\boldsymbol{\zeta}\|_\infty>x)|\to 0.$$

Remark 4. *Proposition 2 indicates that the Kolmogorov distance between the distributions of $\|\Delta\boldsymbol{P}_{\mathcal{S}}\|_\infty$ and $\|\boldsymbol{\zeta}\|_\infty$ converges to zero with rate $O\left((n_0 pd)^{-c}\right)$ for $c>0$ as shown in the proof. However, while increasing d shrinks the distance, without proper adjustment of $n_0 p$, naively changing d will break Assumption 4 and fail sample size requirement as stated in Remark 3.*

The following theorem parallels to the previous proposition except that we replace the population covariance $\boldsymbol{W}^P$ by its plug-in estimator $\widehat{\boldsymbol{W}}^P$. At a high level, as long as the covariance $\boldsymbol{W}^P$ can be estimated well under the supnorm, the Gaussian approximation results remains valid. The proof of Theorem 2 relies on Propositions 1-2.

Theorem 2. *Let $\widehat{\boldsymbol{\zeta}} \sim N(\mathbf{0}, \widehat{\boldsymbol{W}}^P)$ where $\widehat{\boldsymbol{W}}^P$ is the plug-in estimator for $\boldsymbol{W}^P$, with $\widehat{w}_{ij}^P$ following the same form in Equation* (2.17)*, $\widehat{\rho}_{tij}$ given in Equation* (2.9) *and $\widehat{\boldsymbol{U}}_t^*$ given in Equation* (2.24)*, it holds that with probability going to 1,*

$$\sup_{x>0} \left| \mathbb{P}(\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty > x) - \mathbb{P}(\|\widehat{\boldsymbol{\zeta}}\|_\infty > x | \mathcal{D}) \right| \to 0.$$

The above theorem establishes the theoretical foundation for simultaneous testing over multiple graphs. Next we formally state the validity of our testing procedure as well as a power analysis under a general testing $H_{0,\mathcal{S}} : \boldsymbol{\rho}_{ij}^0 = \boldsymbol{c}_{ij}^0,\ \forall (i,j) \in \mathcal{S}$.

Theorem 3. *Under the null, we have that $\mathbb{P}_{H_0}(\Psi_\alpha) \to \alpha$. On the other hand, in the alternative case, if $\max_{(i,j)\in\mathcal{S}} |\frac{\sum_{t=1}^d \xi_{tij}\sqrt{n_t p}(\rho_{tij} - c_{tij})}{\sqrt{d}}| \geq C\sqrt{\log q}\max_{1\leq j\leq r}(w_{jj}^P)^{1/2}$ and C is a large enough constant, then we have $\mathbb{P}_{H_1}(\Psi_\alpha) \to 1$.*

With c_{tij} being zero for all $1 \leq t \leq d$ and $(i,j) \in \mathcal{S}$, it implies that as long as the largest ℓ_1 norm of the partial correction vector $\|\boldsymbol{\rho}_{ij}^0\|_1$ is above the order of $\sqrt{d\log q/(n_o p)}$, the power would be close to 1. Assuming the same order ρ_{tij} across all graphs $1 \leq t \leq d$, i.e., $\rho_{1ij} \asymp \rho_{tij} \asymp \ldots \asymp \rho_{dij}$, the power of the test is close to 1 if $\max_{(i,j)\in\mathcal{S}} |\rho_{1ij}|$ is far larger than $\sqrt{\log q/(dn_o p)}$. In contrast, the corresponding detection boundary becomes $\sqrt{\log q/(n_o p)}$ for a single graph. Therefore, by borrowing the information from multiple graphs/sessions, we are able to reduce the detection accuracy by a factor of root d.

The validity of our c-level test in Equation (2.18) is summarized below before we move to the theoretical properties for Section 2.2.3.

Corollary 1. *Under the null $H'_{0,\mathcal{S}}$, the test is an α level test, i.e., $\mathbb{P}_{H'_0}(\Psi_\alpha) \leq \alpha$. On the other hand, in the alternative case, if $\max_{(i,j)\in\mathcal{S}} \frac{\sum_{t=1}^d \sqrt{n_t p}(|\rho_{tij}| - c)}{\sqrt{d}} \geq C\sqrt{\log q}\max_{1\leq j\leq r}(w_{jj}^P)^{1/2}$ and C is a large enough constant, we have $\mathbb{P}_{H'_1}(\Psi_\alpha) \to 1$.*

In the end, we summarize the estimation bounds under the Frobenius norm for individual temporal covariance and precision matrices obtained in Section 2.2.3. Although exiting results for i.i.d. sample are available in [40], there is no result for correlated samples as derived in our model. We thus provide a self-contained analysis, which might be of independent interest.

Theorem 4. *For our estimation for the temporal covariance matrix $\boldsymbol{U}_t$ and precision matrix $\boldsymbol{A}_t$, we have that for any $\delta > 0$, there exists a constant C which depends on δ and c_e only, such that*

$$\mathbb{P}\Big(\frac{1}{p}\|\widehat{\boldsymbol{U}}_t - \frac{Tr(\boldsymbol{V}_t)}{q}\boldsymbol{U}_t\|_F^2 \geq C\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_2+2}}}\Big) \leq p^{-\delta},$$

$$\mathbb{P}\Big(\frac{1}{p}\|\widehat{\boldsymbol{A}}_t - \frac{q}{Tr(\boldsymbol{V}_t)}\boldsymbol{A}_t\|_F^2 \geq C\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}\Big) \leq p^{-\delta}.$$

Consequently, their scaled Frobenius norms can be consistently estimated, which is sufficient for our main result Theorem 2.

Corollary 2. *For our normalized temporal covariance estimator, we have $\max_{1\leq t\leq d} |\frac{\|\widehat{\boldsymbol{U}}_t^*\|_F^2 - \|\boldsymbol{U}_t\|_F^2}{p}| = o_p(1)$.*

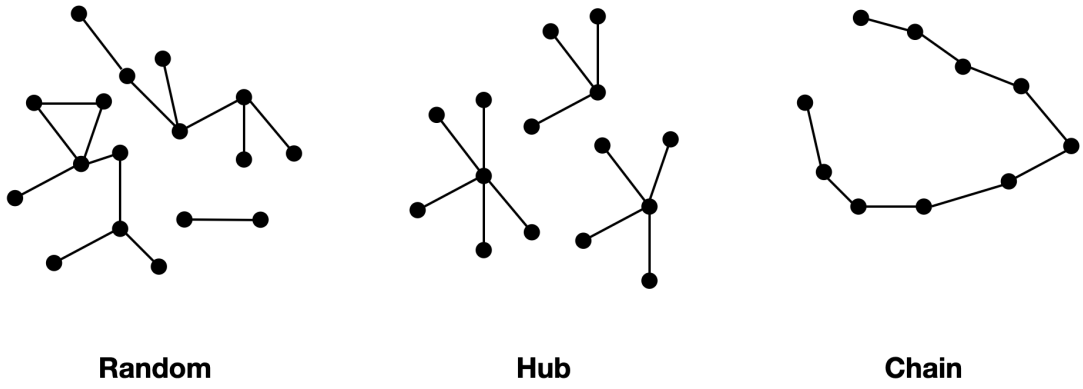

Figure 2.1: Three types of spatial graphs in simulation: random graph, hub graph and band graph.

2.4 Numerical Studies

2.4.1 Simulation Studies

We study the performance of our method via multiple different simulation scenarios. Since simultaneous testing in multiple matrix graphs is largely missing in literature, we demonstrate the effectiveness of our method (M0) via comparison with multiple matrix-variate graph estimation method by [68] (M1) and several multi-graph estimation methods for ordinary Gaussian Graphical Model (M2-M4) , as is shown in Section 2.4.1. We further show that in real-data driven simulation, our method (M0) is still better than the state-of-art method (M1). Finally, in Section 2.4.1, we show that in simultaneous testing, our proposed bootstrap procedure can accurately approximate the $\|\Delta \boldsymbol{P}_{\mathcal{S}}\|$ as theoretically shown in Theorem 2.

We generate our temporal precision matrix $\boldsymbol{A}_t$ following our the parameter space defined in Assumption 6: Recall Equation (2.20), we set $c_{tij} = M(i-j)^{-\alpha_t - 1}$ for $1 \leq j < i \leq q$ with $M = 0.2,\ \alpha_t = 1$, and $\boldsymbol{D}_t$ is simply an identity matrix, for $t \in [d]$. We generate our spatial precision matrix $\boldsymbol{B}_t$ with three different popular structures as in Figure 2.1: a random graph, where the edges between each nodes are randomly generated, with probability of having an edge between i and j as $\sqrt{\frac{3}{q}}$; a hub graph, where all the nodes are divided into several groups, with the number of hub centers as $\lceil \frac{q}{20} \rceil$ and all the rest of nodes divided evenly into each group; a chain graph, which is a special case of banded graph with bandwidth equal to 1. Once the common graph structure is fixed, for each sub-graph t, the strength of the non-zero edge b_{tij} is generated randomly with uniform distribution Unif$(0, 0.3/2^{t-1})$.

As an evaluation, we show the receiver operating characteristic (ROC) curve by varying the test levels for our method. In a typical ROC curve, the y-axis is the true positive rate (TPR), and the x-axis is the false positive rate (FPR), which, in our case, are defined as

$$TPR = \frac{1}{d}\sum_{t=1}^{d} \frac{\sum_{1\leq i<<j\leq q} \mathbb{1}(b_{tij} \neq 0, \widehat{b}_{tij} \neq 0)}{\sum_{1\leq i<<j\leq q} \mathbb{1}(b_{tij} \neq 0)},$$

$$FPR = \frac{1}{d}\sum_{t=1}^{d}\frac{\sum_{1\le i<\le j\le q}\mathbb{1}(b_{tij}=0,\widehat{b}_{tij}\neq 0)}{\sum_{1\le i<\le j\le q}\mathbb{1}(b_{tij}=0)}.$$

Edge-wise Estimation Comparison

We compare our method with several multiple Gaussian Graph estimation methods, which contains the following two categories: (1) the most recent matrix-variate Gaussian multi-graph estimation method proposed by [68] (M1), and (2) multi-graph estimation methods for ordinary Gaussian Graphical Model, such as regression based method by [50] (M2), and optimization based methods by [12] (M3) and [36] (M4). For the baseline methods, the ROC curve can be retrieved by changing tuning parameters, such as the group penalty and sparsity penalty. For methods in the second category, we pre-process the data with whitening over temporal dimension and treat signal at each time point as i.i.d sample. For our method, under different regression tuning parameters, by varying test level α, we get different ROC curves. We discovered that our method is much better than baseline regardless of the tuning parameter, and one example is shown in Figure A.1. The results with $\lambda = 1e-4$ are shown in Figure 2.2. Based on ROC curve, our method recovers the underlying graph structure accurately, and outperforms the baseline methods. Moreover, comparing methods designing for ordinary Gaussian graph, our method is much better when temporal dimension p is large, thanks to the temporal precision estimation based on Cholesky decomposition in Section 2.2.3.

We further modify our method to make precision matrix estimation, which contains two steps: first, by setting a test level α, for an edge (i,j), we can test whether $H_0 : b_{tij} = 0$ for $t \in [d]$ is rejected or not; next, for all edges in precision matrix, based on the test results, we estimate the non-significant edges to be 0, and significant edges following Equation (2.8). Therefore, we can calculate the precision matrix for a particular test level α. The proper choice of α remains to be determined, and we adapt a cross-validation approach. Specifically, we tune the test level α through a grid search of 10 values by estimating the precision matrix on a training set and testing its performance on a validation set through 5-fold cross validation. The training and validation datasets are split with a 80:20 ratio. For l-th fold, denote by our spatial precision matrix estimation based on the training data $\{\widehat{\boldsymbol{B}}_t^{-l}(\alpha)\}_{t=1}^d$, and denote by the sample spatial precision matrix estimation based on the validation data $\{\widehat{\boldsymbol{V}}_t^l\}_{t=1}^d$. To make a fair comparison, we obtain precision matrix estimates of both methods on the same training set, and calculate the loss function on the same validation set, where the loss function is given by

$$l(\alpha) = \sum_{l=1}^{5}\sum_{t=1}^{d}\left\{\log\left[\det\left(\widehat{\boldsymbol{B}}_t^{-l}(\alpha)\right)\right] - \mathrm{tr}\left(\widehat{\boldsymbol{V}}_t^l\widehat{\boldsymbol{B}}_t^{-l}(\alpha)\right)\right\}$$

Of all baseline methods, M1 is the only method that aims for multiple matrix-variate estimation, and recovers the best ROC curve when temporal dimension p is reasonably large. Therefore, we compare our method with M1 especially. We tune group penalty and sparse penalty following the procedure in [68]. Under optimal tuning parameters, to evaluate the accuracy of estimation, we calculate the three different types of loss functions, the matrix 1-norm l_1, the spectral norm l_2 and the Frobenius norm l_F. By fixing number of samples n and temporal dimension p, we vary number of graphs d and spatial dimension q, the comparison between our method M0 and baseline

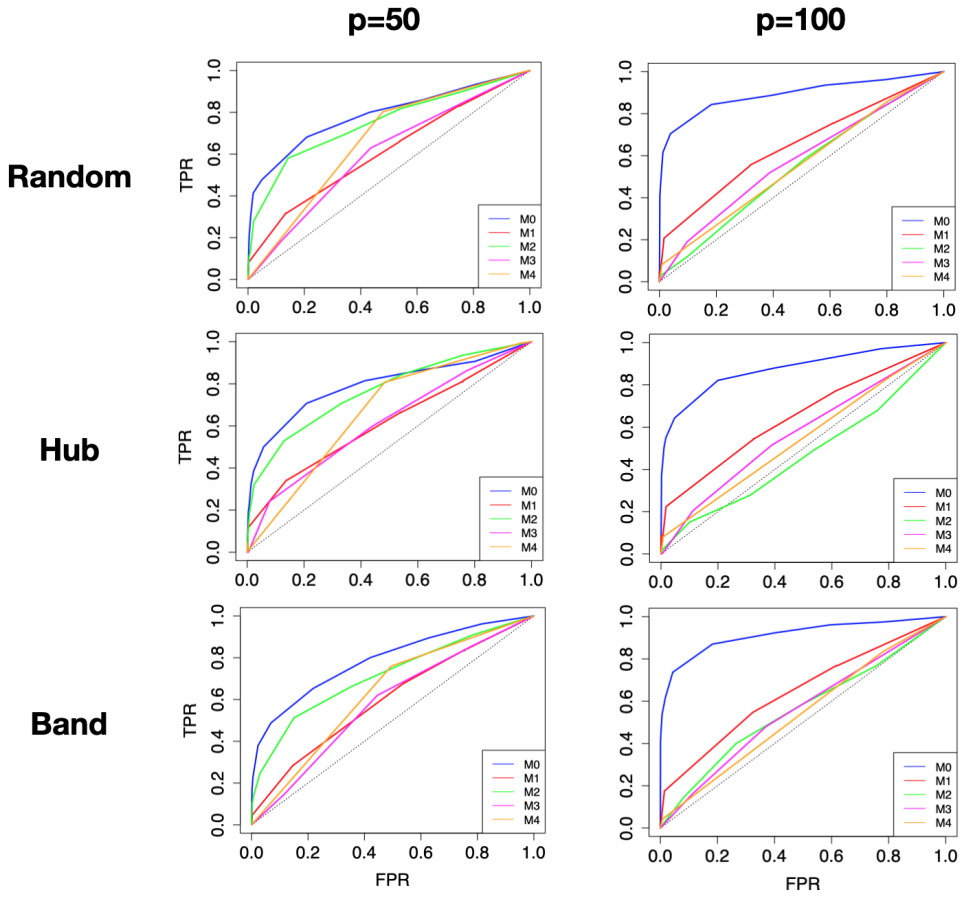

Figure 2.2: Simulation results under different graph configurations and temporal dimensions. We fix $n = 5$, $q = 30$ and $d = 5$. Rows change with types of graphs, and columns correspond to different temporal dimensions. Blue curve corresponds to our method (M0) while other colors correspond to baseline methods. Our method is consistently better than baselines while our advantage is very obvious for large p, thanks to our temporal covariance estimation procedure.

M1 is shown in Table 2.1. We observe that over all settings, M0 outperforms M1 and achieve lower estimation error than M1. Besides, our method is faster in computation than non-convex optimization method, which is slow and not guaranteed to converge.

Simultaneous Test

In this section, we evaluate the performance of propose statistic in Equation (2.12) and verify the correctness of Theorem 2 in finite samples. To set up the simultaneous test, we consider two different index sets: $S_{off} = \{(i, j) : i \neq j\}$ and $S_{zero} = \{(i, j) : b_{tij} = 0,\ \forall t \in d\}$. Due to the limiting computational resources, we consider fixed temporal dimension $p = 50$ and spatial dimension $q = 30$. Then we examine the accuracy of our approximation $\|\widehat{\boldsymbol{\zeta}}^{off}\|_\infty$ and $\|\widehat{\boldsymbol{\zeta}}^{zero}\|_\infty$ to $\|\Delta \boldsymbol{P}_{S_{off}}\|_\infty$ and $\|\Delta \boldsymbol{P}_{S_{zero}}\|_\infty$, where $\widehat{\boldsymbol{\zeta}}^{off}$ and $\widehat{\boldsymbol{\zeta}}^{zero}$ are Gaussian random variables with covariance calculated based on plug-in estimators in Equation (2.17) for each index set. The

Graph	d	q	l_1		l_2		l_F	
			M0	M1	M0	M1	M0	M1
Random	5	30	7.84(0.86)	7.92(1.98)	2.95(0.27)	4.50(0.58)	7.75(0.42)	10.30(0.85)
		50	8.56(1.27)	11.09(0.70)	3.44(0.22)	4.80(0.73)	11.43(0.47)	12.40(1.49)
	10	30	11.05(2.09)	13.47(2.60)	6.12(0.87)	7.82(1.17)	14.90(1.14)	18.58(1.55)
		50	12.07(2.47)	14.63(3.64)	6.51(0.52)	8.21(0.85)	20.47(1.16)	24.96(1.33)
Hub	5	30	10.34(0.85)	16.38(2.12)	3.08(0.23)	5.26(0.73)	6.95(0.28)	8.83(0.96)
		50	11.55(0.51)	17.94(1.03)	3.23(0.10)	5.73(0.23)	8.71(0.26)	11.00(0.50)
	10	30	19.15(2.10)	25.58(3.12)	6.20(0.51)	8.24(0.91)	11.70(1.03)	14.62(1.65)
		50	30.72(3.23)	34.63(4.58)	8.08(0.57)	10.30(1.28)	20.47(0.94)	22.01(1.98)
Band	5	30	4.65(0.12)	6.74(0.13)	2.74(0.06)	3.88(0.08)	7.71(0.09)	10.63(0.12)
		50	9.72(0.20)	10.00(0.10)	3.23(0.06)	3.94(0.07)	11.35(0.10)	13.78(0.13)
	10	30	14.58(0.15)	15.42(0.20)	6.35(0.05)	6.78(0.07)	17.69(0.11)	18.97(0.14)
		50	18.33(0.21)	18.66(0.21)	6.97(0.08)	7.07(0.10)	24.30(0.08)	24.83(0.10)

Table 2.1: Mean and standard deviation of spatial precision estimation for $n = 20, p = 50$.

simulation steps are stated as follows:

- For each type of graph (random, hub, chain), we generate one set of corresponding temporal precision matrix $\{\boldsymbol{A}_t\}_{t=1}^d$ and spatial precision matrices $\{\boldsymbol{B}_t\}_{t=1}^d$.
- Given the precision matrices, for $i = 1, \ldots, 1000$, we generate one data realization, which we can apply our estimation procedure and calculate $\|\Delta \boldsymbol{P}_{S_{off}}\|_\infty$ and $\|\Delta \boldsymbol{P}_{S_{zero}}\|_\infty$. Since their true distributions are unknown, the empirical distributions based on 1000 realizations are denoted by $F_{off}(\cdot)$ and $F_{zero}(\cdot)$.
- For each data realization i, we can apply our simultaneous testing procedure and estimate the $\boldsymbol{W}^P$ corresponding to S_{off} and S_{zero}. By sampling $\widehat{\boldsymbol{\zeta}} \sim N(0, \widehat{\boldsymbol{W}}^P)$, we approximate $\|\Delta \boldsymbol{P}_{S_{off}}\|_\infty$ and $\|\Delta \boldsymbol{P}_{S_{zero}}\|_\infty$ based on the distribution of 3000 bootstrap samples at quantile $\alpha = 0.925$, $\alpha = 0.950$, $\alpha = 0.975$, denoted by $\widehat{q}_{off,\alpha,i}$ and $\widehat{q}_{zero,\alpha,i}$.
- Finally, we can calculate the mean and standard deviation for $\{F_{off}(\widehat{q}_{off,\alpha,i})\}_{i=1}^{1000}$ and $\{F_{zero}(\widehat{q}_{zero,\alpha,i})\}_{i=1}^{1000}$.

The results are shown in Table 2.2 and Table 2.3 for $d = 3$ and $d = 5$, respectively. We observe that the difference between the empirical coverages decreases as the number of sample increases, and the value becomes very small starting from $n = 5$, which demonstrates that our procedure approximates the distribution well and Theorem 2 still holds for finite samples. Comparing Table 2.2 and Table 2.3, we observe that for fixed number of sample, our method performs better for smaller number of graphs, which agrees with our theory that larger number of graphs requires larger number of samples given $\log q$ is small.

n_{sample}	Quantile	Random		Hub		Band	
		S_{off}	S_{zero}	S_{off}	S_{zero}	S_{off}	S_{zero}
5	0.925	0.945(0.006)	0.948(0.004)	0.952(0.003)	0.951(0.004)	0.952(0.006)	0.952(0.007)
	0.95	0.964(0.003)	0.967(0.003)	0.965(0.003)	0.964(0.003)	0.973(0.003)	0.972(0.003)
	0.975	0.979(0.004)	0.981(0.003)	0.987(0.004)	0.985(0.004)	0.985(0.002)	0.985(0.002)
10	0.925	0.931(0.005)	0.933(0.004)	0.915(0.005)	0.917(0.004)	0.930(0.005)	0.925(0.007)
	0.95	0.947(0.003)	0.952(0.003)	0.942(0.004)	0.942(0.004)	0.959(0.004)	0.957(0.004)
	0.975	0.971(0.004)	0.973(0.004)	0.976(0.003)	0.976(0.004)	0.978(0.002)	0.977(0.002)
20	0.925	0.929(0.005)	0.929(0.006)	0.921(0.005)	0.921(0.004)	0.922(0.005)	0.923(0.005)
	0.95	0.950(0.002)	0.951(0.003)	0.949(0.005)	0.948(0.005)	0.950(0.006)	0.950(0.006)
	0.975	0.973(0.003)	0.971(0.004)	0.975(0.001)	0.975(0.001)	0.978(0.004)	0.977(0.004)

Table 2.2: Average of empirical coverages and their standard deviations for $d = 3, p = 50, q = 30$.

n_{sample}	Quantile	Random		Hub		Band	
		S_{off}	S_{zero}	S_{off}	S_{zero}	S_{off}	S_{zero}
5	0.925	0.954(0.005)	0.953(0.006)	0.952(0.005)	0.952(0.005)	0.963(0.004)	0.961(0.004)
	0.95	0.965(0.003)	0.967(0.002)	0.972(0.003)	0.972(0.003)	0.979(0.003)	0.977(0.003)
	0.975	0.982(0.002)	0.983(0.002)	0.990(0.002)	0.989(0.002)	0.995(0.002)	0.994(0.002)
10	0.925	0.936(0.005)	0.939(0.003)	0.945(0.005)	0.942(0.005)	0.927(0.004)	0.930(0.003)
	0.95	0.958(0.004)	0.958(0.004)	0.968(0.004)	0.965(0.002)	0.947(0.004)	0.947(0.004)
	0.975	0.973(0.002)	0.976(0.001)	0.979(0.001)	0.979(0.001)	0.972(0.004)	0.972(0.003)
20	0.925	0.930(0.005)	0.931(0.007)	0.918(0.005)	0.917(0.006)	0.926(0.006)	0.927(0.006)
	0.95	0.947(0.005)	0.949(0.005)	0.944(0.004)	0.944(0.004)	0.954(0.004)	0.956(0.003)
	0.975	0.976(0.004)	0.976(0.004)	0.973(0.004)	0.974(0.004)	0.981(0.002)	0.982(0.003)

Table 2.3: Average of empirical coverages and their standard deviations for $d = 5, p = 50, q = 30$.

Experimental Data Driven Simulation

The details of the experiment can be found in Section 2.4.2. We apply our method to V4 LFPs during the cue stage, and estimate the spatial and temporal precision matrix. For spatial precision, we pick alpha level $\alpha = 0.01$ with Bonferroni correction to generate a sparse spatial precision matrix. With these spatio-temporal estimation, we re-generate simulation data, which share the same dimension as real data with $n = 1000$, $p = 50$, $q = 96$. We set the number of sessions $d = 3$ since M1 is slow on high-dimensional data. As is shown in Figure 2.3, our method (M0) outperforms the multiple matrix graph estimation method (M1) regardless of tuning parameter. We also notice that under this setting with a reasonable lambda, we almost recover ROC curve perfectly. While our data-driven procedure prefers larger tuning parameter to guarantee the correctness of theory, we further notice that in practice, the optimal lambda can be determined by cross-validation.

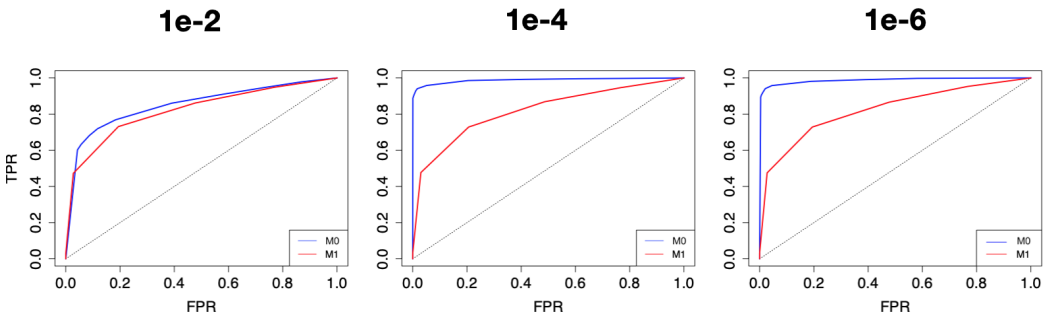

Figure 2.3: Simulation results using real data estimates under different tuning parameter λ values. The dataset is of the same dimension as real data, $n = 1000$, $p = 50$, $q = 96$, but only keep $d = 3$ for fast computation of M1. Our method (M0) is always better than M1 regardless of tuning parameter. As the number of sample is large, M2 is similar to our method, thus the curve is omitted here.

Since our real data has a relatively smaller dimension, $q = 96$, the difference between our method (M0) and modified inference method (M2) is not obvious especially when number of sample is large. By setting $n = 10$, we compare the ROC curve between our method (M0)

and M2, which is shown in Figure 2.4. Our method (M0) is better than M2 when processing high-dimensional data with small sample size.

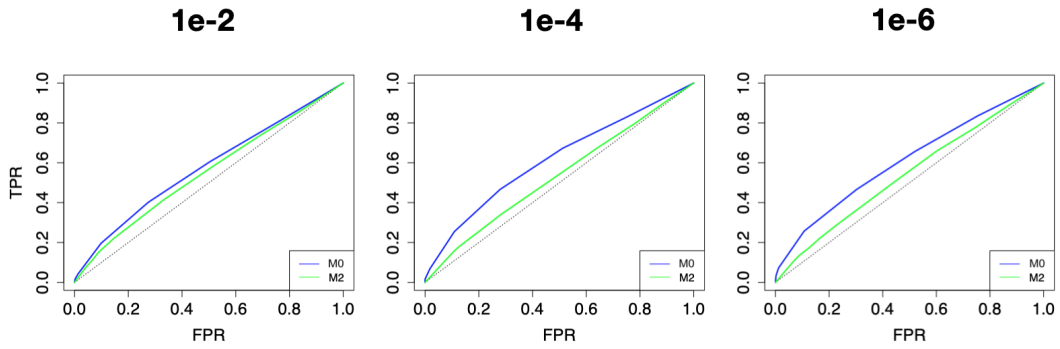

Figure 2.4: Simulation results using real data estimates under different tuning parameters. The dataset is of the same spatial and temporal dimension as real data, but only keep $n = 10$ to make it a high dimensional inference problem. Our method (M0) is always better than M2 regardless of tuning parameter.

2.4.2 Experimental Data Analysis

The details of the data can be found in Section 1.1, and we are focused on the complete dataset which contains 5 different experimental sessions. Before applying our method to the data, we need to verify that for a pair of edge (i, j), the sign of ρ_{tij} remains the same for all $1 \leq t \leq d$. We calculate the sample partial correlation estimate $\widehat{\boldsymbol{\rho}}_t^{samp}$ for each sub-graph separately, and plot the edge values in sub-graph against each other in Figure 2.5. We also provide the sample estimate $\widehat{\boldsymbol{B}}_t^{samp}$ for precision matrix in Figure A.2 in the appendix. We observe that across different sessions, the sample estimates for a fixed edge are strongly correlated (correlation coefficient > 0.94 for all sessions pairs in Figure 2.5), and the signs are mostly the same, thus our assumption holds.

We further define four different experimental stages: fixation stage (200ms), cue stage (50ms), early delay stage (the first 250ms of delay stage) and late delay stage (the last 250ms of delay stage). We leave 100 trials as cross-validation data for tuning parameter λ in group Lasso following similar procedure as in Section 2.4.1.

Within-area Inference

We apply our method in PFC and V4 separately, each with 96 electrodes located in a 10×10 utah array. Since the spatial location for electrodes is known, we are particularly interested in inferring the relationship between neural connectivity and physical distance. Notice the significance of the edge is encoded in our test statistic, which can also be interpreted as sum of partial correlation for each edge over multiple graphs. Therefore, we can calculate the average test statistic over all edges for a particular physical distance, which will reflect how the connectivity is related to edge distance in physical space. As in Figure 2.6, we discover that the test statistic declines as the

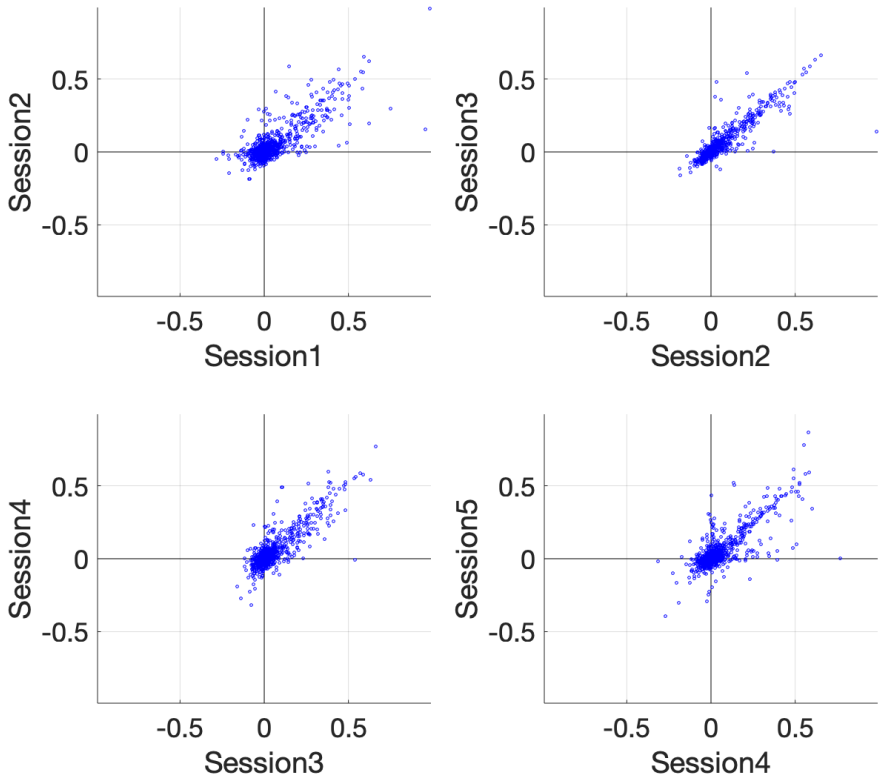

Figure 2.5: Sample estimate of $\boldsymbol{\rho}_t$ for each session plotting against each other. For each panel, we plot vec($\widehat{\boldsymbol{\rho}}_t^{samp}$) vs. vec($\widehat{\boldsymbol{\rho}}_{t+1}^{samp}$), for $1 \leq t \leq 4$. Each blue dot corresponds to an edge value in partial correlation matrix. The sample estimate shows that edges in each sub-graph are strongly correlated, and the signs keep the same for most edges, especially the edges with strong connectivity.

physical distance increases for all experimental stages in both area, which echos with the fact that the correlation of neuron activity depends strongly on physical distances from previous studies [28, 60]. This serves as a side proof that our test statistic is intepretable and reasonable in real data.

Next, for each area and each experimental stage, we apply our method on the corresponding data segments. Define the connectivity strength of node i as $\frac{1}{d}\sum_{j:(i,j)\in S_*}\sum_{t=1}^{d}|\rho_{tij}|$, where S_* is the set of significant edges at level $\alpha = 0.05$ for each stage, then we can investigate the distribution of connectivity strength changing with area and experimental stages. In Figure 2.7, we observe that the within area connectivity are strongest during fixation and cue stage, while it declines during delay stage. We also observe that PFC seems to be more connected than V4. Especially, at the delay stage, when animal needs to proceed the visual signals, connectivity in both PFC and V4 decreases; the connectivity in V4 dies down quickly while connectivity in PFC remains on a high level during delay stage. To further test our observation, we define the change of connectivity for node i as $\frac{1}{d}\sum_{j:(i,j)\in S'_*}\sum_{t=1}^{d}(|\rho^1_{tij}| - |\rho^2_{tij}|)$, where ρ^1_{tij} is the partial correlation between i and

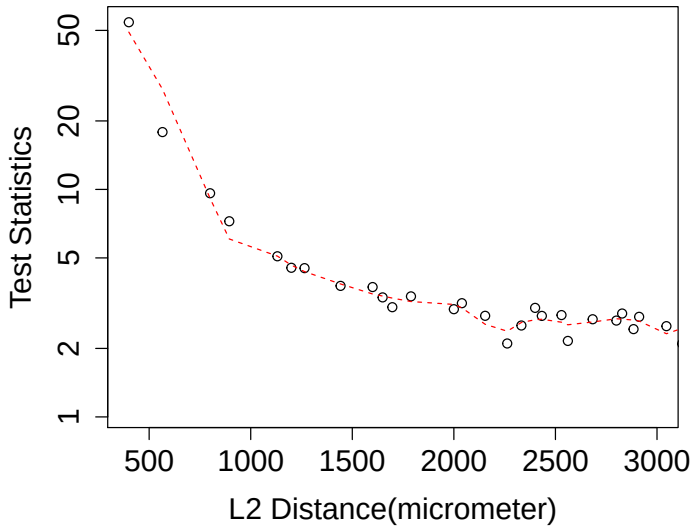

Figure 2.6: The average test statistic vs physical distance during late delay period in V4. Notice that the test statistic declines as the physical distance increase. This phenomenon is consistently identified over all experimental stages, both in PFC and V4

j for graph t at late delay stage, ρ^2_{tij} is the partial correlation between i and j for graph t at cue stage, S'_* is the set of significantly changed edges between cue stage and late delay stage at level $\alpha = 0.05$. The spatial distribution of change of connectivity is shown in Figure A.3. We observe negative changes in connectivity strength for both PFC and V4 when switching from cue stage to delay stage.

Finally, as the cue appears at one of eight random directions, we analyze the connectivity distribution with respect to each area, time stage and cue type. For V4, during fixation, there are no significant differences in connectivity respect to each cue. However, starting from the cue stage, we observe that in V4, the connectivity peaks at $225°$. The connectivity figures for eight different cues are shown in Figure 2.8. This gives evidence that V4 neurons do have a preferred direction, not only in terms of firing rate/tuning curve of single neuron, but also in terms of the connectivity strength which is defined by sum of partial correlation of bulk activity. To confirm our finding with evidence from statistical test, for V4 at cue stage, we collect one dataset containing 210 trials with cue appearing at $225°$, and the other dataset containing same number of trials with cue uniformly appearing at 7 other direction. We test if the change of connectivity is significant at level $\alpha = 0.05$, and plot the change of connectivity strength in Figure A.4, following similar procedure in Figure A.3. Indeed, we see a striking increase in connectivity and verified that V4 gains a stronger connectivity for cue at $225°$. For PFC, there is no clear sign of a *preferred* direction, which is reasonable as it's not a visual area as V4. However, when compare the figure of PFC during cue stage (Figure A.5) with the one during late delay stage (Figure A.6), we observe that PFC seems to gradually "shift" to attend stimulus at angle $225°$ as well. In Figure A.5, cues at $45°/135°/180°$ all show stronger connectivity than $225°$; however, in Figure A.6, $225°$ seems to be the strongest. This indicates that the cross communication between PFC and V4 leads PFC

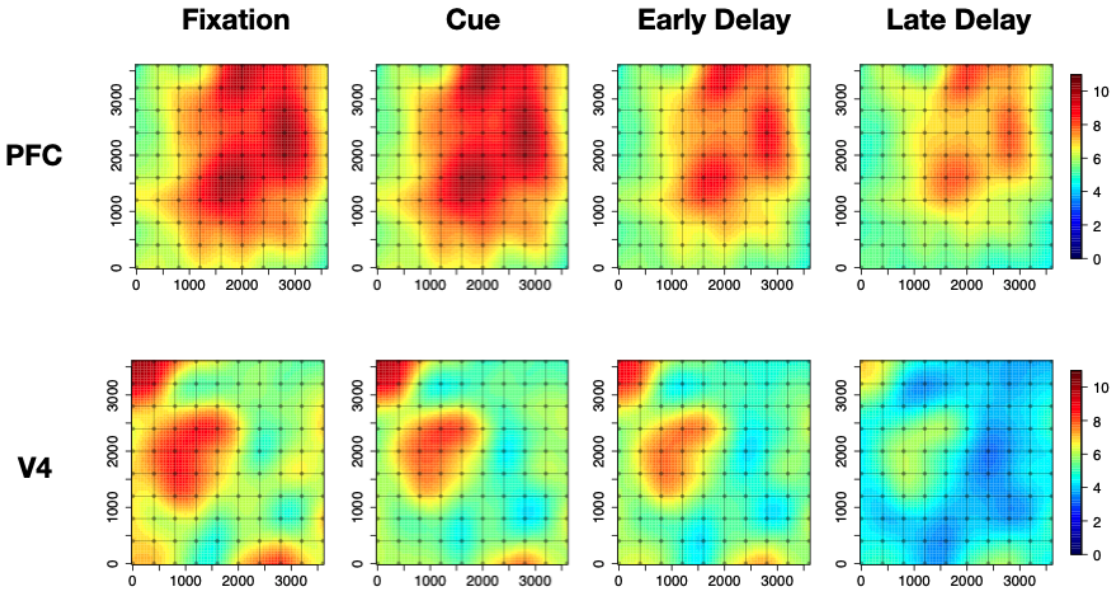

Figure 2.7: Connectivity strength distribution over 2D array for PFC and V4 over various experimental stages. The connectivity in both area decays during the delay stage, but V4 seems to be more influenced and less active than PFC.

to respond similarly as V4, but a more careful inference needs to be carried out later.

Cross-area Inference

The declining connectivity for within area drives us to investigate the cross-area connectivity. Now combining the electrodes in both area and applying our procedure, we are capable to identify significant cross-area edges. We sub-sample the electrodes in each area by taking every other node along the physical dimension, making it $5 \times 5 = 25$ electrodes for each region, and 625 cross-area edges in total. The electrode subsampling can be justified by the fact that adjacent nodes are usually strongly connected, thus skipping one neighbour will not negatively impact our results, but ensures numerical stability. Besides, it saves computational resources and make visualization of cross-region edges cleaner. Testing at $\alpha = 0.05$, the significant cross-area edges are shown in Figure 2.9. We identify least number of edges during cue stage, when the animal is supposed to observe the cue. The cross-area edges seem to be recovering during early delay stage, and reach to the same number as fixation stage. At late delay stage, when the animal is about to make a choice, the most number of cross-edges are identified.

To characterize the overall cross-region connectivity, we implemented the simultaneous test for the set of cross-area edges at test level $\alpha = 0.05$. In Table 2.4, we show the test results at different c-levels. Specifically, we identify that the two area are most strongly connected during late delay period, and there is few connectivity between two areas during cue period. The cross-area connectivity appears to be on the same level for fixation stage and early delay stage.

Combining with the results from within-area inference, we identify that cross-area connectivity

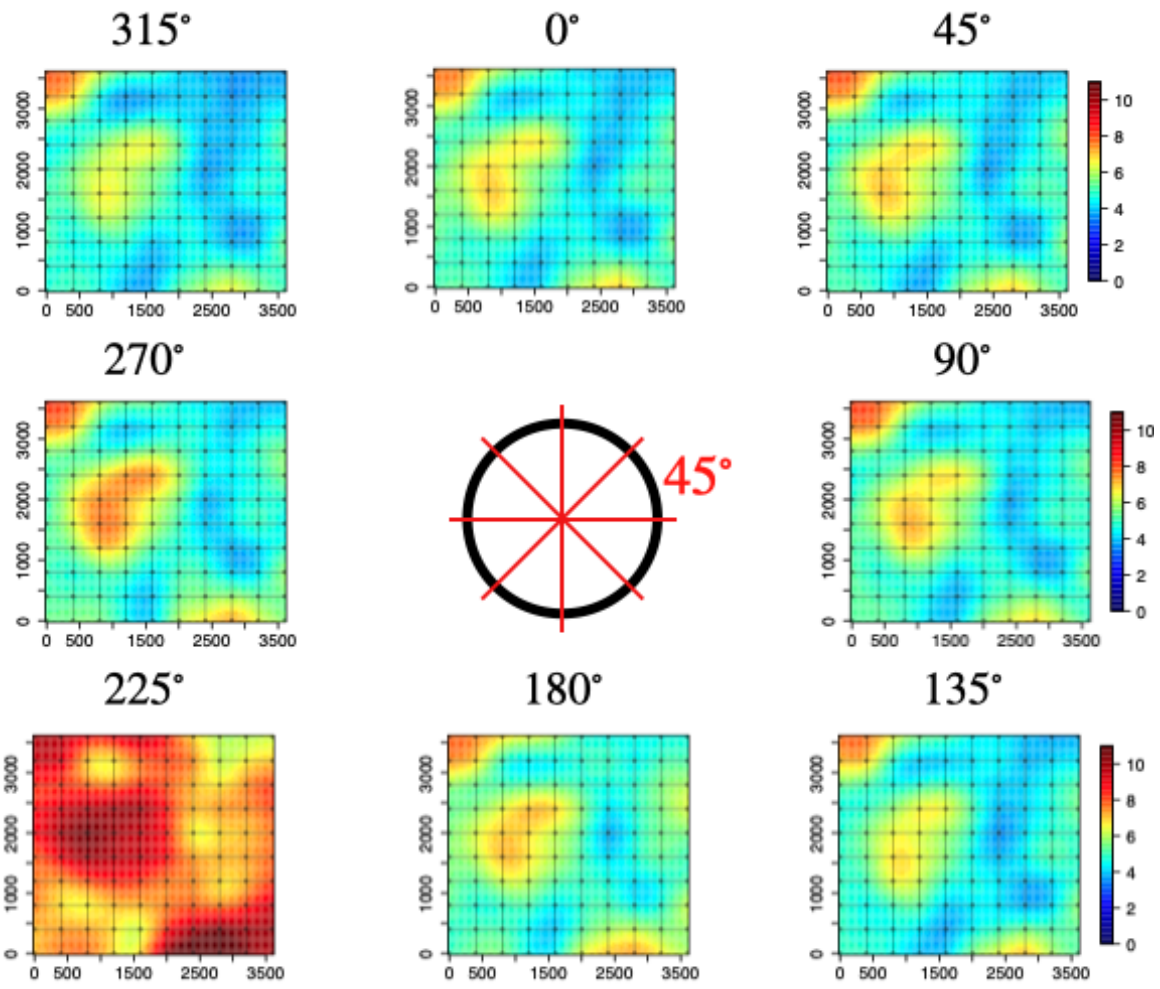

Figure 2.8: Connectivity strength distribution over 2D array for V4 for eight different cues during cue stage. Eight different panel corresponds to eight different cues appearing at different angles. The connectivity at 225° *shows a much stronger amplitude than other directions. Similar figure for PFC is shown in appendix.*

gets enhanced during the delay stage, when the animal has to process the visual signal and prepare to make a choice, while within-area connectivity is suppressed. On the other hand, during cue stage, while the animal is focused on seeing the cue, both areas seem to function on its own and there is little communication. The early delay stage seems to be a transition stage, during which the within-area connectivity decays while cross-area connectivity recovers to the level of fixation stage and continues to increase. These observations echo with previous studies that neural variability in the spiking of neurons declines during the stimulus onset [19], and visual stimuli causes a substantial decrease in correlation of cortical neurons [57]. We also discover that PFC is more active than V4 especially during delay stage, while V4 is mainly involved during fixation and cue stage. As is shown by [35], robust sustained activity in PFC is found during delay stage, which is further supported by our result. We also identify that the bulk neuron activity in V4

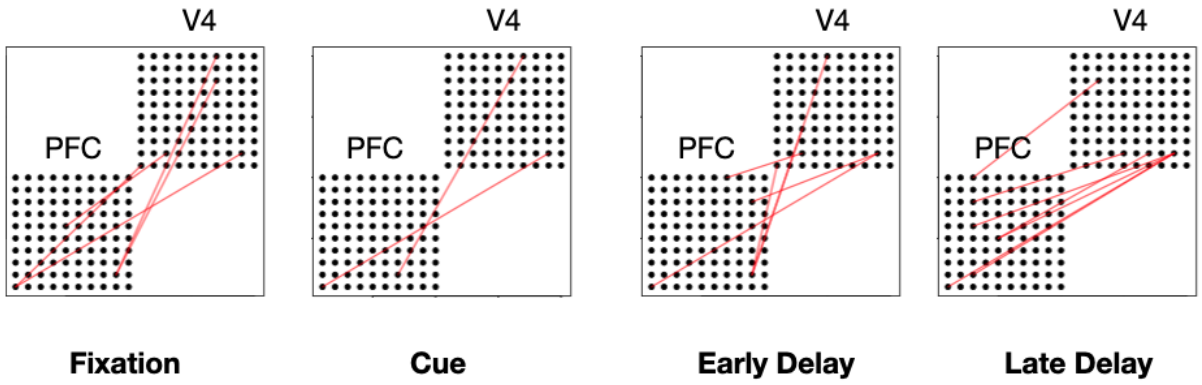

Figure 2.9: Significant cross-region edges for PFC and V4 over various experimental stages. X and Y axis are electrode numbering along each dimension. Lower left shows electrodes in PFC, while upper right shows electrodes in V4. Red lines are the significant cross-area edges.

shows a preferred direction for cue at $225°$.

c-level	Fixation	Cue	Early Delay	Late Delay
0	*		*	*
0.01	*		*	*
0.02				*

*Table 2.4: Simultaneous test results for cross-area edges at different c-levels with test level $\alpha = 0.05$. Entry with * represents significant test result.*

2.5 Conclusion

In this paper, we propose a linear-functional based test using partial correlation estimator to detect sparse edges and infer existence and strength of connectivity between two groups of nodes in multiple matrix-variate Gaussian Graphical Models. The spatial dimension, temporal dimension and number of graphs are allowed to diverge and even exceed the number of samples.

Both our model and our assumptions are driven by the practical concerns in neural data analysis. In real data, we observe the within-area connectivity and cross-area connectivity changes accordingly, as the animal entered different experimental stages. Especially, within-area connectivity peaks during early experimental stages, while cross-are connectivity grows when the animal processes the visual signal during late delay stage. Our inference results are illuminating for scientists to understand the activity and connectivity of PFC and V4 during visual tasks.

Our method is the first attempt to address the simultaneous c-level test problem in multiple matrix-variate Gaussian graphs. It would be interesting to extend our method to other popular yet non-Gaussian type of graphs such as Poisson networks. Besides, we currently implemented group Lasso for our regression model which involves one tuning parameter; in the future, a tuning-free or scale-free method such as self-tuned Dantzig selector and scaled Lasso is desirable to handle the issue of heterogeneity and correlation in regression with data from multiple matrix-variate

Gaussian graphical models. These directions are beyond the scope of this work and will be interesting for future directions.

Discussion on Significant Result

Throughout Figure 2.6 - Figure 2.8, we draw our conclusions based on significance test. However, non-significant results may due to either weak signals or large uncertainty. In our case, the analytical form in Equation 2.17 with Assumption 3 guarantees that the variance of all edges are asymptotically on the same level, thus specific edges are unlikely to be dropped from our graphs due solely to the magnitude of uncertainty.

To investigate this further, we analyzed the variance of the test statistic for all the figures we generated, and discovered that the variance is on a constant level indeed for all the figures we have in the thesis. Two examples, corresponding to line plot as Figure 2.6 and heatmap plot as Figure 2.7, are shown here. In Figure 2.10, we show the averaged standard deviation of our test statistic for each physical distance, and we observe no dramatic change over the physical distance. In Figure 2.11, we calculate the square root of averaged variance of test statistic for each node i ($\sqrt{\frac{1}{q}\sum_{j=1}^{q}|\text{Var}(P_{(i,j)})|}$), and show the heatmap. For each stage and each area, the spatial distribution of the variance is uniform across all spatial nodes, which indicates that spatial distribution of connectivity is indeed determined by our signal strength, and is not heavily influenced by the variance of test statistic.

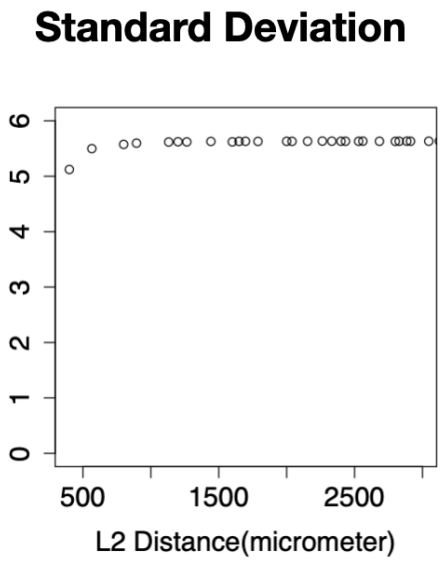

Figure 2.10: The averaged variance of test statistic vs physical distance during late delay period in V4. Notice that different from test statistic, the variance seems to be stable over the physical distance, thus ruling out the possibility of having larger variance for larger physical distance.

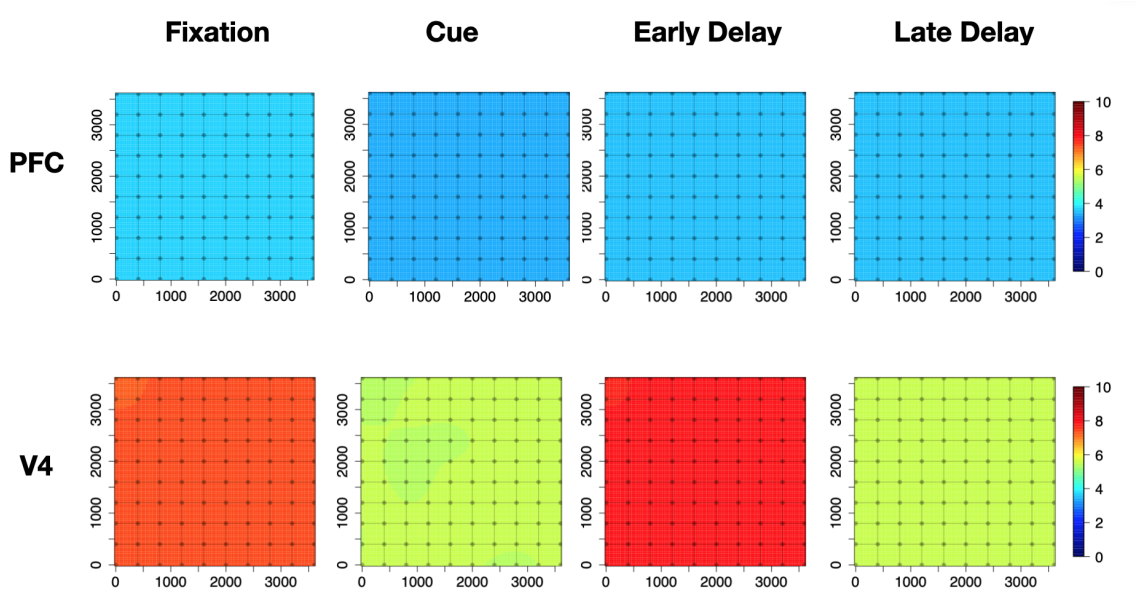

Figure 2.11: The square root of averaged variance of test statistic distribution over 2D array for PFC and V4 over various experimental stages. There are no obvious spatial patterns for all areas and experimental stages.

Chapter 3

Latent Dynamic Factor Analysis of High-Dimensional Neural Recordings

This chapter is taken from work submitted to *NeurIPS 2020, the Thirty-fourth Conference on Neural Information Processing Systems* aside from minor changes for style consistency. I collaborated with co-author Heejong Bong, Zhao Ren, Matthew A. Smith, Valerie Ventura, and Robert E. Kass.

In this chapter, we would like to investigate **Q3**, that is, how strong the connectivity is across area. From the study in Chapter 2, imposing one single matrix-variate graph to multiple brain areas might be too rigid, as 1) there might be multiple factors and it's inappropriate to assume the autocorrelation to be the same over all spatial nodes, as is discussion in Section A.5.1, and 2) we would like to investigate bi-directional dynamic cross-area connectivity (the lead-lag dependency) which changes over time and characterizes the information flow in both directions.

Therefore, we introduce a factor model to estimate dynamic spatio-temporal dependence between PFC and V4 from LFP data. By adding a noise component to CCA and extending it into multiple components, our model can be interpreted as a general model of dynamic CCA; by special design for factor covariance and noise covariance, our model can also be viewed as a cross-region generalization of factor models such as GPFA. We assume that the observed data consists of two parts, (i) the factor part and (ii) the noise part, as in a typical factor model, and the novelties of our model comes from two folds: (i) For the factor part, motivated by CCA, we design a joint latent factor which combines latent components from all brain areas and directly samples from a full covariance matrix. We further notice that, while having a full covariance for joint factors enables more flexibility, we have large number of parameters to be estimated. To handle this situation, we impose the constraint that each in-region and cross-region sub-precision matrix block are banded. Furthermore, even for the non-zero entries, an ℓ_1 penalty is added to impose sparsity constraint. (ii) For the noise part, we assume that the auto-covariance of noise is the Kronecker product of spatial covariance and temporal covariance, which is a common assumption in Matrix-variate Graphical Model designed for processing spatio-temporal data [67]. The spatial precision matrix can be estimated with Graphical Lasso to impose sparsity structure [26], while the temporal precision matrix can be estimated with modified Cholesky decomposition to impose banded structure[5, 40]. Finally, by putting all parts together, we optimize the model parameters with Expectation–Maximization (EM) algorithm, and the parameter tuning is based on K-fold

cross-validation.

3.1 Latent Dynamic Factor Analysis of High-dimensional Time Series

Our notations are slightly different than the notations in previous two chapters, since we assume different factors in different brain areas. We denote the area by the superscript k, i.e., $\boldsymbol{X}^1$ is the observation data from the first area. As a consequence, we denote the i-th sample in region k by $\boldsymbol{X}^k[i]$. For practical reason, we only consider two areas in model description, while our model can be easily extended to three or more areas.

We treat the case of two groups of time series observed, repeatedly, N times. Let $\boldsymbol{X}^1_{t,\cdot} \in \mathbb{R}^{p_1}$ and $\boldsymbol{X}^2_{t,\cdot} \in \mathbb{R}^{p_2}$ be p_1 and p_2 recordings at time t in each of the two groups, for $t = 1, \ldots, T$. As in [65], we assume that a q-dimensional latent factor $\boldsymbol{Z}^k_{t,\cdot} \in \mathbb{R}^q$ drives each group, here, each brain region, according to the linear relationship

$$\boldsymbol{X}^k_{t,\cdot} \mid \boldsymbol{Z}^k_{t,\cdot} = \boldsymbol{\mu}^k_{t,\cdot} + \boldsymbol{\beta}^k \cdot \boldsymbol{Z}^k_{t,\cdot} + \boldsymbol{\epsilon}^k_{t,\cdot}, \tag{3.1}$$

for brain region $k = 1, 2$, where $\boldsymbol{\mu}^k_{t,\cdot} \in \mathbb{R}^{p_k}$ are mean vectors, $\boldsymbol{\beta}^k \in \mathbb{R}^{p_k \times q}$ are matrices of constant factor loadings, and $\boldsymbol{\epsilon}^k_{t,\cdot} \in \mathbb{R}^{p_k}$ are errors centered at zero (independently of the latent vectors Z). We are interested in the pairwise cross-group dependencies of the latent vectors $\boldsymbol{Z}^1_{\cdot,f}$ and $\boldsymbol{Z}^2_{\cdot,f}$, for $f = 1, \ldots, q$. As in our related work [6], we assume that the time series of these latent vectors follows a multivariate normal distribution

$$\begin{pmatrix} \boldsymbol{Z}^1_{\cdot,f} \\ \boldsymbol{Z}^2_{\cdot,f} \end{pmatrix} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_f), \quad f = 1, \ldots, q, \tag{3.2}$$

where $\boldsymbol{\Sigma}_f$ describes all of their simultaneous and lagged dependencies, both within and between the two vectors. We assume the N sets of random vectors $(\boldsymbol{\epsilon}, \boldsymbol{Z})$ are independent and identically distributed. We let $\boldsymbol{P}_f$ be the correlation matrix corresponding to $\boldsymbol{\Sigma_f}$, and write its inverse as

$$\boldsymbol{\Pi}_f = \boldsymbol{P}_f^{-1} = \left(\begin{array}{c|c} \boldsymbol{\Pi}_f^{11} & \boldsymbol{\Pi}_f^{12} \\ \hline \boldsymbol{\Pi}_f^{12\top} & \boldsymbol{\Pi}_f^{22} \end{array} \right) \tag{3.3}$$

where $\boldsymbol{\Pi}_f^{11}$ and $\boldsymbol{\Pi}_f^{22}$ are the scaled auto-precision matrices and $\boldsymbol{\Pi}_f^{12}$ is the scaled cross-precision matrix. We now assume finite-range partial autocorrelation and cross-correlation for $(\boldsymbol{Z}^1_{t,f}, \boldsymbol{Z}^2_{t,f})$, so that $\boldsymbol{\Pi}_f^{11}$, $\boldsymbol{\Pi}_f^{22}$ and $\boldsymbol{\Pi}_f^{12}$ in Equation (3.3) have a banded structure. Specifically, for $k, l = 1, 2$, we assume there is a value h_f^{kl} such that $\boldsymbol{\Pi}_f^{kl}$ is a $(2h_f^{kl} + 1)$-diagonal matrix. Because our goal is to address the cross-region connectivity and lead-lag relationship, we are particularly interested in the estimation of $\boldsymbol{\Pi}_f^{12}$ for each latent factor $f = 1, \ldots, q$. Note that the non-zero elements $\Pi^{12}_{f,(t,s)}$ determine associations between the latent pair $\boldsymbol{Z}^1_{\cdot,f}$ and $\boldsymbol{Z}^2_{\cdot,f}$, which are simultaneous when $t = s$ and lagged when $t \neq s$. Finally, we model the noise in Eq. (3.1) as a Gaussian random vector

$$(\boldsymbol{\epsilon}^k_{1,\cdot}; \boldsymbol{\epsilon}^k_{2,\cdot}; \ldots; \boldsymbol{\epsilon}^k_{t,\cdot}) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Phi}^k), \quad k = 1, 2, \tag{3.4}$$

where we allow $\mathbf{\Phi}^k$ to have non-zero off-diagonal elements to account for within-group spatiotemporal dependence. We assume Φ^k can be written in Kronecker product form

$$\mathbf{\Phi}^k = \mathbf{\Phi}^k_{\mathcal{T}} \otimes \mathbf{\Phi}^k_{\mathcal{S}},\ k = 1, 2, \tag{3.5}$$

where $\mathbf{\Phi}^k_{\mathcal{S}}$ and $\mathbf{\Phi}^k_{\mathcal{T}}$ are the spatial and temporal components of Φ^k, as is often assumed for spatiotemporal matrix-normal distributions, e.g., [22]. Although this is a strong approximation, implying, for instance, that the auto-correlation of every $\boldsymbol{X}^k_{\cdot,i}$ is proportional to $\mathbf{\Phi}^k_{\mathcal{T}}$, we regard $\mathbf{\Phi}_k$ as a nuisance parameter: our primary interest is $\mathbf{\Sigma}_f$ in Eq. (3.2). We also assume an auto-regressive process of order at most h^k_ϵ, so that $\mathbf{\Gamma}^k_{\mathcal{T}} = \left(\mathbf{\Phi}^k_{\mathcal{T}}\right)^{-1}$ is a $(2h^k_\epsilon + 1)$-diagonal matrix. In our simulation we show that we can recover $\mathbf{\Sigma}_f$ accurately even when the Kronecker product and bandedness assumptions fail to hold.

The model in Equations (1)-(5) generalizes other known models. First, when $q = 1$, and $\boldsymbol{Z}^1 = \boldsymbol{Z}^2$ remains constant over time, in the noiseless case ($\boldsymbol{\epsilon}_k = 0$), it reduces to the probabilistic CCA model of [3]. Thus, model (1)-(5) can be viewed as a denoising, multi-level and dynamic version of probabilistic CCA. Second, when $k = 1$, the Gaussian processes are stationary, and the $\boldsymbol{\epsilon}$ vectors are white noise, (1)-(5) reduces to GPFA [65]. Thus, (1)-(5) is a two-group, nonstationary extension of GPFA that allows for within-group spatio-temporal dependence.

3.1.1 Identifiability and Sparsity Constraints

Despite the structure imposed on $\mathbf{\Phi}_k$ in Eq. (3.5), parameter identifiability issues remain. Our model in Eqs. (3.1), (3.2) and (3.4) induces the marginal distribution of the observed data $(\boldsymbol{X}^1, \boldsymbol{X}^2)$:

$$\left(\boldsymbol{X}^1_{1,\cdot}; \boldsymbol{X}^1_{2,\cdot}; \ldots; \boldsymbol{X}^2_{t,\cdot}\right) \sim \mathrm{N}\left((\boldsymbol{\mu}^1_{1,\cdot}; \boldsymbol{\mu}^1_{2,\cdot}; \ldots; \boldsymbol{\mu}^2_{t,\cdot}), \boldsymbol{S}\right) \tag{3.6}$$

where $\boldsymbol{S}$ is the marginal covariance matrix given by:

$$\boldsymbol{S} = \begin{bmatrix} \mathbf{\Phi}^1_{\mathcal{T}} \otimes \mathbf{\Phi}^2_{\mathcal{S}} & 0 \\ 0 & \mathbf{\Phi}^2_{\mathcal{T}} \otimes \mathbf{\Phi}^2_{\mathcal{S}} \end{bmatrix} + \sum_{f=1}^{q} \begin{bmatrix} \mathbf{\Sigma}^{11}_f \otimes (\boldsymbol{\beta}^1_f \boldsymbol{\beta}^{1\top}_f) & \mathbf{\Sigma}^{12}_f \otimes (\boldsymbol{\beta}^1_f \boldsymbol{\beta}^{2\top}_f) \\ \mathbf{\Sigma}^{12\top}_f \otimes (\boldsymbol{\beta}^2_f \boldsymbol{\beta}^{1\top}_f) & \mathbf{\Sigma}^{22}_f \otimes (\boldsymbol{\beta}^2_f \boldsymbol{\beta}^{2\top}_f) \end{bmatrix}. \tag{3.7}$$

The family of parameters

$$\boldsymbol{\theta}^{\{\alpha^1, \alpha^2\}} = \left\{ \begin{array}{l} \mathbf{\Sigma}^{\{\alpha^1_1, \alpha^2_1\}}_1, \ldots, \mathbf{\Sigma}^{\{\alpha^1_q, \alpha^2_q\}}_q,\ \mathbf{\Phi}^1_{\mathcal{S}} - \sum_{f=1}^{q} \alpha^1_f \boldsymbol{\beta}^1_f \boldsymbol{\beta}^{1\top}_f,\ \mathbf{\Phi}^2_{\mathcal{S}} - \sum_{f=1}^{q} \alpha^2_f \boldsymbol{\beta}^2_f \boldsymbol{\beta}^{2\top}_f, \\ \mathbf{\Phi}^1_{\mathcal{T}},\ \mathbf{\Phi}^2_{\mathcal{T}},\ \boldsymbol{\beta}^1,\ \boldsymbol{\beta}^2,\ \boldsymbol{\mu}^1,\ \boldsymbol{\mu}^2 \end{array} \right\}, \tag{3.8}$$

where $\mathbf{\Sigma}^{\{\alpha^1_f, \alpha^2_f\}}_f = \left\{ \mathbf{\Sigma}_f + \begin{bmatrix} \alpha^1_f \mathbf{\Phi}^1_{\mathcal{T}} & 0 \\ 0 & \alpha^2_f \mathbf{\Phi}^2_{\mathcal{T}} \end{bmatrix} \right\}$, induce the same marginal distribution in Eq. (3.6), for all $\alpha^1, \alpha^2 \in \mathbb{R}^q$ (notice that $\boldsymbol{\theta} = \boldsymbol{\theta}^{\{\mathbf{0},\mathbf{0}\}} = \{\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_q,\ \mathbf{\Phi}^1_{\mathcal{S}},\ \mathbf{\Phi}^2_{\mathcal{S}},\ \mathbf{\Phi}^1_{\mathcal{T}},\ \mathbf{\Phi}^2_{\mathcal{T}},\ \boldsymbol{\beta}^1,\ \boldsymbol{\beta}^2,\ \boldsymbol{\mu}^1,\ \boldsymbol{\mu}^2\}$ is the original parameter). Preliminary analysis of LFP data indicated that strong cross-region dependence occurs relatively rarely. We therefore resolve this lack of identifiability by choosing the solution given by maximizing the likelihood with an L1 penalty, under the assumption that the inverse cross-correlation matrix Π^{12}_f is a sparse $(2h^{12}_f + 1)$-diagonal matrix.

3.1.2 Latent Dynamic Factor Analysis of High-dimensional Time Series (LDFA-H)

Given N simultaneously recorded pairs of neural time series $\{\boldsymbol{X}^1[n], \boldsymbol{X}^2[n]\}_{n=1,\dots,N}$, the maximum penalized likelihood estimator (MPLE) of the inverse correlation matrix of the latent variables solves

$$\begin{aligned}\left(\widehat{\boldsymbol{\Pi}}_1, \dots, \widehat{\boldsymbol{\Pi}}_q\right) = \operatorname{argmin} \; & -\frac{1}{N}\sum_{n=1}^{N} l\left(\boldsymbol{\theta}; \boldsymbol{X}^1[n], \boldsymbol{X}^2[n]\right) + \sum_{f=1}^{q}\sum_{k,l=1}^{2} \left\|\boldsymbol{\Lambda}_f^{kl} \odot \boldsymbol{\Pi}_f^{kl}\right\|_1 \\ & \text{s.t. } \boldsymbol{\Gamma}_{\mathcal{T}}^k \text{ is } (2h_\epsilon^k + 1)\text{-diagonal},\end{aligned} \tag{3.9}$$

where the log-likelihood is

$$l\left(\boldsymbol{\theta}; \boldsymbol{X}^1, \boldsymbol{X}^2\right) = -\log\det \boldsymbol{S} - (\boldsymbol{X}^1_{1,\cdot} - \boldsymbol{\mu}^1_{1,\cdot}; \dots; \boldsymbol{X}^2_{t,\cdot} - \boldsymbol{\mu}^2_{t,\cdot})^\top \boldsymbol{S}^{-1} (\boldsymbol{X}^1_{1,\cdot} - \boldsymbol{\mu}^1_{1,\cdot}; \dots; \boldsymbol{X}^2_{t,\cdot} - \boldsymbol{\mu}^2_{t,\cdot}), \tag{3.10}$$

with $\boldsymbol{S}$ defined in Eq. (3.7), and the constraints are

$$\Lambda^{kl}_{f,(t,s)} = \begin{cases} \infty, & (t,s) : |t-s| > h_f^{kl}, \\ \lambda_f, & (t,s) : 0 < |t-s| \le h_f^{kl}, \;\; k \neq l, \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

for factor $f = 1, \dots, q$ and brain region $k = 1, 2$. The first constraint forces the corresponding $\Pi^{kl}_{f,(t,s)}$ to zero and thus imposes a banded structure for Π_f^{kl}, and the second assigns the same sparsity constraint λ_f on the off-diagonal elements of Π_f^{12}. Finally, to make calibration of tuning parameters computationally feasible, we use the same bandwidth to latent precision and noise precision within a region and to latent precision across regions, respectively, and the same sparsity parameter:

$$h_f^{kk} = h_\epsilon^k = h_{\text{auto}}, \;\; h_f^{12} = h_{\text{cross}} \;\text{ and }\; \lambda_f = \lambda_{\text{cross}},$$

for each factor $f = 1, \dots, q$ and region $k = 1, 2$.

The remaining hyperparameters become the temporal bandwidths h_{auto} and h_{cross}; the sparsity penalty λ_{cross}; and the number of latent factors q. The bandwidths are chosen using domain knowledge and preliminary data analyses. We determine the remaining parameters by 5-fold cross-validation (CV).

Solving Eq. (3.9) requires $\boldsymbol{S}^{-1}$. Because it is not available analytically and a numerical approximation is computationally prohibitive, we solve Eq. (3.9) using an EM algorithm [24]. Let $\boldsymbol{\theta}^{(r)}$ be the parameter estimate at the r-th iteration. We consider the data $\{\boldsymbol{X}^1[n], \boldsymbol{X}^2[n]\}_{n=1,\dots,N}$ to be incomplete observations of $\{\boldsymbol{X}^1[n], \boldsymbol{Z}^1[n], \boldsymbol{X}^2[n], \boldsymbol{Z}^2[n]\}_{n=1,\dots,N}$. In the E-step, we estimate the conditional mean and covariance matrix of each $\{\boldsymbol{Z}^1[n], \boldsymbol{Z}^2[n]\}$ with respect to $\{\boldsymbol{X}^1[n], \boldsymbol{X}^2[n]\}$ and $\boldsymbol{\theta}^{(r)}$. Given these sufficient statistics, the MPLE decomposes into two separate minimizations of

1. the negative log-likelihood of $\boldsymbol{\Sigma}_f$, w.r.t. the latent factor model (Eq. (3.2)) and
2. the negative log-likelihood of $\boldsymbol{\Phi}^1_{\mathcal{S}}$, $\boldsymbol{\Phi}^2_{\mathcal{S}}$, $\boldsymbol{\Phi}^1_{\mathcal{T}}$, $\boldsymbol{\Phi}^2_{\mathcal{T}}$, $\boldsymbol{\beta}^1$, $\boldsymbol{\beta}^2$, $\boldsymbol{\mu}^1$, $\boldsymbol{\mu}^2$ w.r.t. the observation model (Eqs. (3.1) and (3.4)).

With the noise correlation and latent factor correlation disentangled, the M-step reduces to easy sub-problems. For example, the minimization with respect to $\boldsymbol{\Sigma}_f$ is a graphical Lasso problem [26] and the minimization with respect to $\boldsymbol{\Phi}_{\mathcal{S}}^k$ and $\boldsymbol{\Phi}_{\mathcal{T}}^k$ is a maximum likelihood estimation of a matrix-variate distribution [22]. We thus obtain an affordable M-step, and alternating E and M-steps produces a solution to the MPLE problem. We derive the full formulations in Appendix B.1. Code is provided at `https://github.com/AutoAnonymous/ldfa_anon`.

3.2 Results

One major novelty of our method is its accounting of auto-correlated noise in neural time series to better estimate cross-region associations in CCA type analysis. This is illustrated in Section 3.2.1 based on simulated data. Then in Section 3.2.2, we apply LDFA-H to experimental data to examine the lead-lag relationships across two brain areas and the spatial distribution of factor loadings.

3.2.1 Real-data Based Simulation

We simulated $N = 1000$ i.i.d. neural time series $\boldsymbol{X}^k$ of duration $T = 50$ from Eq. (3.1) for brain regions $k = 1, 2$. The latent time series $\boldsymbol{Z}^k$ were generated from Eq. (3.2) with $q = 1$ pair of factors and correlation matrix $\boldsymbol{P}_1$ depicted in Fig. 3.1(a). The noise $\boldsymbol{\epsilon}^k$ was taken to be the $N = 1000$ trials of the experimental data analyzed in Section 3.2.2, first permuted to remove cross-region correlations, then contaminated with white noise to modulate the strength of noise correlation relative to cross-region correlations. The resulting temporal noise correlation matrices, found by averaging correlations over all pairs of simulated time series, are shown in Fig. 3.1(b), for four levels of white noise contamination. The magnitudes of cross-region correlation and within-region noise auto-correlation are quantified by the determinant of each matrix, known as the generalized variance [54]; their logarithms are provided atop the panels in Fig. 3.1(a) and Fig. 3.1(b). Generalized variance ranges from 0 (identical signals) to 1 (independent signals). Other simulation details are in Appendix B.

We note that the simulation does not satisfy some of the model assumptions in Section 3.1. The noise vectors $\boldsymbol{\epsilon}^k$ are not matrix-variate distributed as in Eqs. (3.4) and (3.5) and the derived $\boldsymbol{\Gamma}_{\mathcal{T}}^k$ does not satisfy a banded structure as in Eq. (3.9). Also, the latent auto-correlations (Fig. 3.1) are not banded as assumed in Eq. (3.9).

We applied LDFA-H with $q = 1$ factor, $h_{\text{cross}} = 10$, h_{auto} equal to the maximum order of the auto-correlations in the 2000 observed simulated time series, and λ_{cross} determined by 5-fold CV. Fig. 3.2 shows LDFA-H cross-precision matrix estimates corresponding to the four level of noise correlation shown in Fig. 3.1(b). They closely match the true Π_1^{12} shown in the right most panel of Fig. 3.1(a).

We also applied five other methods to estimate cross-region connections in the simulated data. They include the popular averaged pairwise correlation (APC); correlation of averaged signals (CAS); and CCA [30], applied to the NT observed pairs of multivariate random vectors $\{\boldsymbol{X}_{t,\cdot}^1, \boldsymbol{X}_{t,\cdot}^2\}_{n,t\in[N]\times[T]}$ to estimate the cross-correlation matrix between the canonical variables; as well as DKCCA [52] and Method A ([6]). The first four methods do not explicitly provide

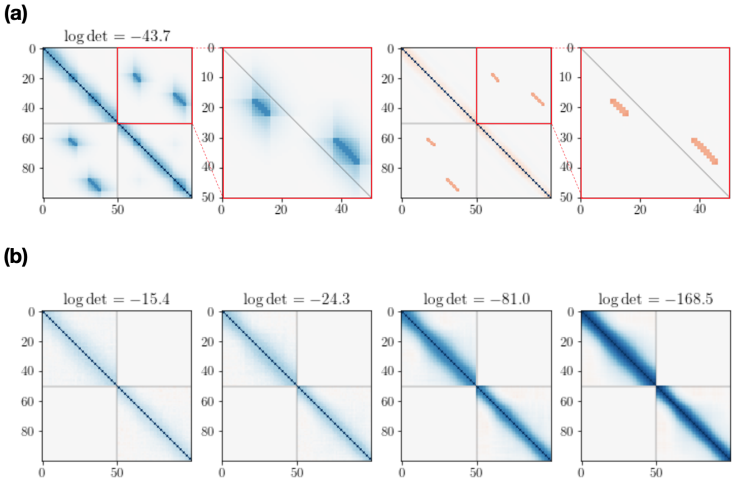

Figure 3.1: Simulation settings. (a) True correlation matrix P_1 for latent factors $\boldsymbol{Z}^1_{\cdot,1}$ and $\boldsymbol{Z}^2_{\cdot,1}$ from model in Eq. (3.2)*; close-up of the cross-correlation matrix; corresponding precision matrix $\boldsymbol{\Pi}_1 = \boldsymbol{P}_1^{-1}$; and close-up of cross-precision matrix $\boldsymbol{\Pi}_1^{12}$ (Eq.* (3.3)*). Matrix axes represent the duration, $T = 50$ ms, of the time series. Factors $\boldsymbol{Z}^1$ and $\boldsymbol{Z}^2$ are associated in two epochs: $\boldsymbol{Z}^2$ precedes $\boldsymbol{Z}^1$ by 7ms from $t = 13$ to 19ms, and $\boldsymbol{Z}^1$ precedes $\boldsymbol{Z}^2$ by 7ms from $t = 33$ to 42ms. (b) Noise auto-correlation matrices (Eq.* (3.5)*) for pairs of simulated time series at four strength levels.* $\log\det$ *in (a) and (b) measure correlation strengths.*

cross-precision matrix estimates, so we display their cross-correlation matrix estimates in Fig. 3.3, along with LDFA-H cross-correlation estimates in the last row. It is clear that only LDFA-H successfully recovered the true cross-correlations shown in the second panel of Fig. 3.1(a), at all auto-correlated noise levels.

3.2.2 Experimental Data Analysis from Monkey Saccade Task

We now report the analysis of LFP data in areas PFC and V4 of a monkey during an eye saccade task. One trial of the experiment consisted of four stages: (i) fixation: the animal fixated at the center of the screen; (ii) cue: a cue appeared on the screen randomly at one of eight locations; (iii) delay: the animal had to remember the cue location while maintaining eye fixation; (iv) choice: the monkey made a saccade to the remembered cue location. We focused our analysis on the 500 ms delay period, when the animal both processed cue information and prepared a saccade. LFP data were recorded for $N = 1000$ trials by two 96-electrode Utah arrays each implanted in PFC and V4, β band-passed filtered, and down-sampled from 1 kHz to 100 Hz.

We applied LDFA-H using $h_{\text{auto}} = h_{\text{cross}} = 10$, corresponding to 100 ms (at 100 Hz); the LFP β-power envelops have frequencies between $12.5Hz$ to $30Hz$, and $h_{\text{auto}} = 10$ just enables the slowest filtered signal to complete one full oscillation period. The other tuning parameters were determined by 5-fold CV over $\lambda_{\text{cross}} \in \{0.0001, 0.001, 0.01, 0.1\}$ and $q \in \{5, 10, 15, 20, 25, 30\}$, yielding optimal values $\lambda_{\text{cross}} = 0.01$ and $q = 10$. The fitted factors were ranked based on the Frobenius norms of their covariance matrices $\|\boldsymbol{\Sigma}_f\|_F^2$; norms are plotted versus f in decreasing

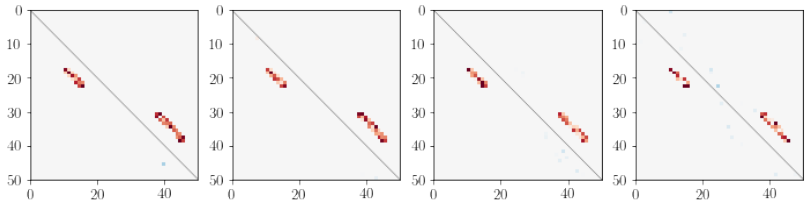

Figure 3.2: Simulation results: LDFA-H cross-precision matrix estimates. Estimates of Π_1^{12}*, shown in the right-most panel of Fig. 3.1(a), using LDFA-H, for the four noise auto-correlation strengths shown in Fig. 3.1(b). LDFA-H identified the true cross-area connections at all noise strengths.*

order in Fig. B.1, and $\log_{10} \|\mathbf{\Sigma}_f\|_F^2$ of the four most dominant factors are provided atop each panel in Fig. 3.4(a). Factor loadings (slightly smoothed over space) for the 96 V4 electrodes are shown in Fig. 3.4(a) for the top four factors (first four columns of the estimate of β^k in Eq. (3.9), with area $k = 1$ being V4), arranged spatially according to electrode positions on the Utah array. The factors have different spatial modes over the physical space of the Utah array. For example, the dominant first factor has positive weights concentrated along a vertical strip on the left of the array, especially in the mid-to-upper left, and negative weights along a vertical strip to the right, separated by roughly 2000 microns.

We also summarized, for each factor f, the temporal information flow at time t from V4 to PFC and to V4 from PFC with $I_{f,out}(t) = \sum_{t'>t} \left|\widehat{\mathbf{\Pi}}^{12}_{f,(t,t')}\right|$ and $I_{f,in}(t) = -\sum_{t'<t} \left|\widehat{\mathbf{\Pi}}^{12}_{f,(t,t')}\right|$, respectively, where $\widehat{\mathbf{\Pi}}_f$ is the inverse correlation matrix estimate in Eq. (3.9). Figure 3.4(b) displays smoothed $I_{f,out}(t)$ and $I_{f,in}(t)$ as functions of $t \in [100, 400]ms$ for the top four factors. Lead-lag relationships between V4 and PFC change dynamically over time, and the information flow tends to peak either at the beginning of the delay period, when the animal must remember the cue, or at the end, when it must make a saccade decision. We observe the strongest information flow from V4 to PFC in the dominant first factor, when the animal needs to process the visual signal from V4 during the delay period. We also observe different information flow patterns: (1) asymmetric (first factor): information in and out avoid conflicting with each other and peak at different times; (2) parallel (second factor): information in and out are parallel and oscillate over time; (3) symmetric (third and fourth factors): information in and out are symmetric, meaning they are activated and suppressed simultaneously.

3.3 Conclusion

To identify dynamic interactions across brain regions we have developed LDFA-H, a nonstationary, multi-group extension of GPFA that allows for within-group spatio-temporal dependence among high-dimensional neural recordings. Although we treated the two-group case, and applied it to interactions across two brain regions, several groups can be handled with obvious, and straightforward modifications. The approach could, in principle, be applied to many different types of time series, but it has some special features: first, like all methods based on sparsity, it assumes a small number of large effects are of primary interest; second, it uses repetitions,

here, repeated trials, to identify time-varying dependence; third, because the within-group spatio-temporal structure is not of interest, the method can remain useful even with some modest within-group model misspecification.

We applied LDFA-H to LFP data, while GPFA has been applied mainly to neural spike count data. In the analysis of spike counts, we have been struck by the strong attenuation of effects due to Poisson-like noise, as discussed in [60] and references therein. A version of LDFA-H built for Poisson-like counts, or for point processes, could be the subject of additional research. It may also be advantageous to model spatial dependence explicitly, perhaps based on physical distance between electrodes, analogously to what was done in [60], and there may be important simplifications available in the temporal structure as well. In addition, it would be helpful to include statistical inferences for assessing effects. In the future, we hope to pursue these possible directions, and refine the application of this promising approach to the analysis of high-dimensional neural data.

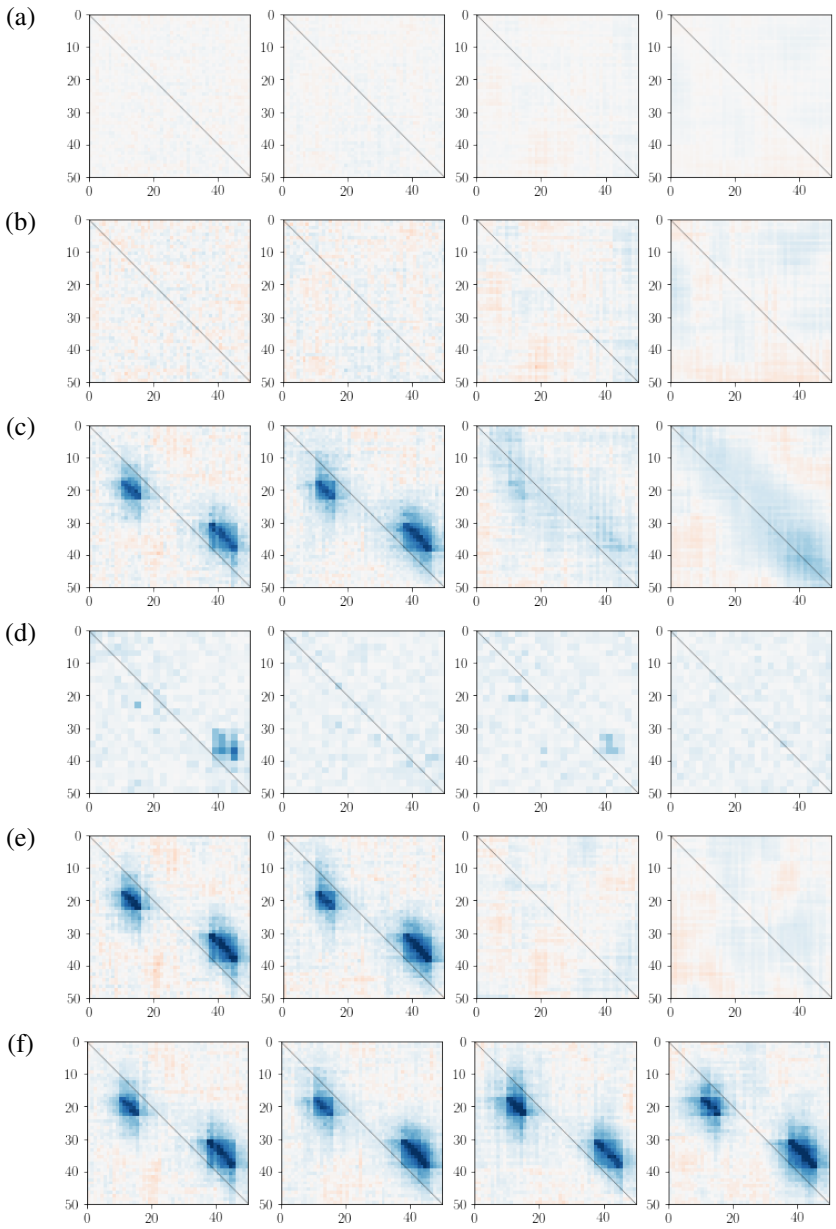

Figure 3.3: Simulation results: cross-correlation matrix estimates. Estimates of Σ_1^{12} using (a) averaged pairwise correlation (APC), (b) correlation of averaged signal (CAS), (c) canonical correlation analysis (CCA, [30]), (d) dynamic kernel CCA (DKCCA, [52]), (e) Method A ([6]), and (f) LDFA-H under four noise correlation levels. Only LDFA-H successfully recovered the true cross-correlation at all noise auto-correlation strengths.

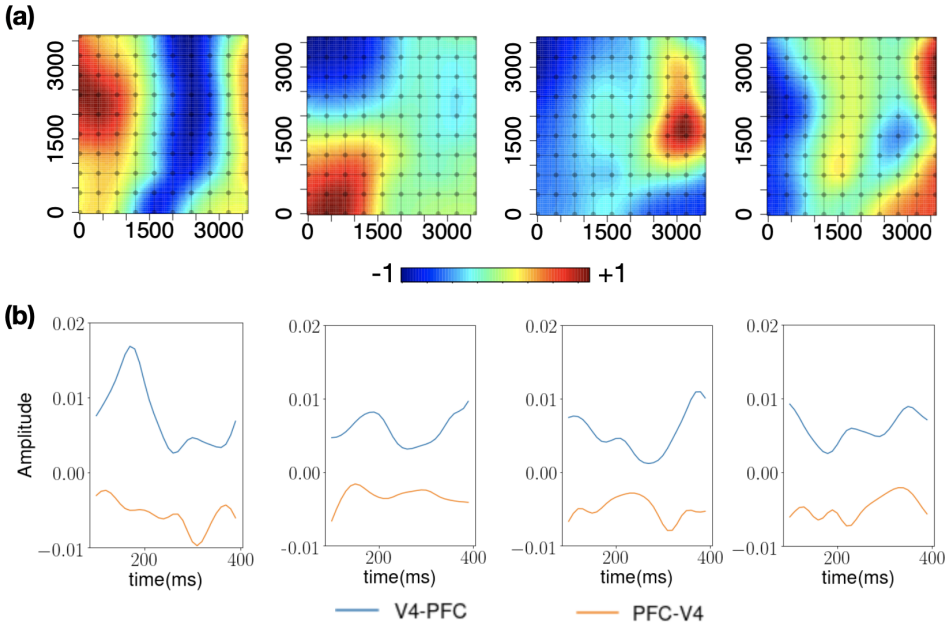

Figure 3.4: Experimental data results for the top 4 factors. (a) Factor loadings, rescaled between -1 and 1, plotted against the electrode coordinates (μm) of the V4 Utah array. Factors have different spatial modes over the physical space of the Utah array. $\log_{10} \|\Sigma_f\|_F^2$*, written atop the panels, measures the strength of each factor. Notice that the strength of the first factor is over 100 order larger than the second largest factor. (b) Dynamic information flow from* $V4 \rightarrow PFC$ *(blue) and* $PFC \rightarrow V4$ *(orange). In and out flows seem to peak either at the beginning or at the end of the delay period, and different couplings of the two flows may indicate different communication modes between V4 and PFC.*

Bibliography

[1] Allen, G. I. and Tibshirani, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *The Annals of Applied Statistics*, 4(2):764–790. 1.3

[2] Armington, J. C. and Bloom, M. B. (1974). Relations between the amplitudes of spontaneous saccades and visual responses. *Journal of Optical Society of America*, 64(9):1263–1271. 2.2.3

[3] Bach, F. R. and Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, Berkeley, CA. 3.1

[4] Bichot, N. P., Rossi, A. F., and Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area v4. *Science*, 308(5721):529–534. 1.1

[5] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227. 1.4, 1.5, 2, 2.2.3, 2.3, 3, B.1

[6] Bong, H., Ventura, V., Smith, M., and Kass, R. (2020a). Discovering lead-lag relationship between two brain areas. *arXiv preprint*. (document), 3.1, 3.2.1, 3.3

[7] Bong, H., Ventura, V., Smith, M., and Kass, R. (2020b). Discovering lead-lag relationship between two brain areas. *arXiv preprint*. 1.5, B.1

[8] Brincat, S. L. and Miller, E. K. (2015). Frequency-specific hippocampal-prefrontal interactions during associative learning. *Nature Neuroscience*, 18(4):576–581. 2.2.3

[9] Brincat, S. L. and Miller, E. K. (2016). Prefrontal cortex networks shift from external to internal modes during learning. *Journal of Neuroscience*, 36(37):9739–9754. 2.2.3

[10] Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., and Bressler, S. L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences*, 101(26):9849–9854. 1.5

[11] Buzsáki, G. (2004). Large-scale recording of neuronal ensembles. *Nature Neuroscience*, 7(5):446–451. 1

[12] Cai, T. T., Li, H., Liu, W., and Xie, J. (2016a). Joint Estimation of Multiple High-dimensional Precision Matrices. *Statistica Sinica*, 26(2):445–464. 1.4, 2.4.1

[13] Cai, T. T., Ren, Z., and Zhou, H. H. (2016b). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59. 2.3

[14] Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block

thresholding. *The Annals of Statistics*, 40(4):2014–2042. A.2.1, A.4

[15] Chang, J., Qiu, Y., Yao, Q., and Zou, T. (2018). Confidence regions for entries of a large precision matrix. *Journal of Econometrics*, 206(1):57 – 82. 1.3, 1

[16] Chen, X. and Liu, W. (2018). Graph Estimation for Matrix-variate Gaussian Data. *Statistica Sinica*, 29:479–504. 1.3, 1.4, 2.2.1, 3

[17] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819. 1, 1.4, 2, 2.2.2, 2.3, A.3, A.4

[18] Chernozhukov, V., Chetverikov, D., and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, 162(1):47–70. A.3

[19] Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., Bradley, D. C., Smith, M. A., Kohn, A., Movshon, J. A., Armstrong, K. M., Moore, T., Chang, S. W., Snyder, L. H., Lisberger, S. G., Priebe, N. J., Finn, I. M., Ferster, D., Ryu, S. I., Santhanam, G., Sahani, M., and Shenoy, K. V. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, 13(3):369–378. 2.4.2

[20] Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397. 1.4

[21] Dandekar, S., Privitera, C., Carney, T., and Klein, S. A. (2012). Neural saccadic response estimation during natural viewing. *Journal of Neurophysiology*, 107(6):1776–1790. PMID: 22170971. 2.2.3

[22] Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274. 3.1, 3.1.2

[23] de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574. 2

[24] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. 3.1.2

[25] Fornito, A., Zalesky, A., and Breakspear, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage*, 80:426 – 444. Mapping the Connectome. 1.3

[26] Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441. 1.3, 3, 3.1.2

[27] Gautier, E. and Tsybakov, A. B. (2013). *Pivotal Estimation in High-Dimensional Regression via Linear Programming*, pages 195–204. Springer Berlin Heidelberg, Berlin, Heidelberg. A.5, A.5, A.5

[28] Goris, R. L. T., Movshon, J. A., and Simoncelli, E. P. (2014). Partitioning neuronal variability.

Nature Neuroscience, 17(6):858–865. 2.4.2

[29] Herreras, O. (2016). Local Field Potentials: Myths and Misunderstandings. *Frontiers in Neural Circuits*, 10:101. 1

[30] Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer. (document), 3.2.1, 3.3

[31] Janková, J. and van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229. 1.3

[32] Johnston, R., Snyder, A. C., Khanna, S. B., Issar, D., and Smith, M. A. (2020). The eyes reflect an internal cognitive state embedded in the population activity of cortical neurons. *bioRxiv*. 1.1

[33] Kelly, R. C., Smith, M. A., Kass, R. E., and Lee, T. S. (2010). Local field potentials indicate network state and account for neuronal response variability. *Journal of Computational Neuroscience*, 29(3):567–579. 1

[34] Lauritzen, S. L. (1996). *Graphical Models*. Oxford Univ. Press. 1.3

[35] Leavitt, M. L., Mendoza-Halliday, D., and Martinez-Trujillo, J. C. (2017). Sustained activity encoding working memories: Not fully distributed. *Trends in Neurosciences*, 40(6):328 – 346. 2.4.2

[36] Lee, W. and Liu, Y. (2015). Joint Estimation of Multiple Precision Matrices with Common Structures. *Journal of Machine Learning Research*, 16:1035–1062. 1.4, 2.4.1

[37] Leng, C. and Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, 99(3):703–716. 1.4

[38] Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978. 1, 1.3, 2.2.1, 2.2.1

[39] Liu, W.-D., Lin, Z., and Shao, Q.-M. (2008). The asymptotic distribution and berry–esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *The Annals of Applied Probability*, 18(6):2337–2366. 1.4, 2

[40] Liu, Y. and Ren, Z. (2017). Minimax Estimation of Large Precision Matrices with Bandable Cholesky Factor. *arXiv preprint*, page arXiv:1712.09483. 2.2.3, 2.2.3, 2.3, 2.3, 3, A.2.1, A.4

[41] Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., and Benali, H. (2006). Partial correlation for functional brain interactivity investigation in functional mri. *NeuroImage*, 32(1):228 – 237. 2

[42] Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322. B.1

[43] McKee, Matthew (2019). A braingate electrode array with a dime for size comparison. [Online; accessed Jan 16, 2020]. (document), 1.1

[44] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462. 1.3, 2.2.1

[45] Mitzdorf, U. (1987). Properties of the evoked potential generators: Current source-density

analysis of visually evoked potentials in the cat cortex. *International Journal of Neuroscience*, 33(1-2):33–59. 1

[46] Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–784. 1.1

[47] Ossandón, J. P., Helo, A. V., Montefusco-Siegmund, R., and Maldonado, P. E. (2010). Superposition model predicts eeg occipital activity during free viewing of natural scenes. *Journal of Neuroscience*, 30(13):4787–4795. 2.2.3

[48] Peterson, C., Stingo, F. C., and Vannucci, M. (2015). Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174. PMID: 26078481. 1.4

[49] Qiu, Y. and Zhou, X.-H. (2018). Estimating c-level partial correlation graphs with application to brain imaging. *Biostatistics*, pages 1–18. kxy076. 1.3, 2, 2

[50] Ren, Z., Kang, Y., Fan, Y., and Lv, J. (2019). Tuning-free heterogeneous inference in massive networks. *Journal of the American Statistical Association*, 114(528):1908–1925. 1.4, 2, 2.4.1, A.2.1, A.3

[51] Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026. 1, 1.3

[52] Rodu, J., Klein, N., Brincat, S. L., Miller, E. K., and Kass, R. E. (2018). Detecting multivariate cross-correlation between brain regions. *Journal of Neurophysiology*, 120(4):1962–1972. PMID: 29947591. (document), 1.5, 3.2.1, 3.3

[53] Sapountzis, P. and Gregoriou, G. (2017). Neural signatures of attention: Insights from decoding population activity patterns. *Frontiers In Bioscience*, 22:2069–2090. (document), 1.1

[54] Sengupta, A. (2004). Generalized variance. *Encyclopedia of statistical sciences*. 3.2.1

[55] Seth, A. K., Barrett, A. B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, 35(8):3293–3297. 1.4, 1.5, 2, 2.2.3

[56] Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., Groblewski, P. A., Ahmed, R., Arkhipov, A., Bernard, A., Billeh, Y. N., Brown, D., Buice, M. A., Cain, N., Caldejon, S., Casal, L., Cho, A., Chvilicek, M., Cox, T. C., Dai, K., Denman, D. J., de Vries, S. E. J., Dietzman, R., Esposito, L., Farrell, C., Feng, D., Galbraith, J., Garrett, M., Gelfand, E. C., Hancock, N., Harris, J. A., Howard, R., Hu, B., Hytnen, R., Iyer, R., Jessett, E., Johnson, K., Kato, I., Kiggins, J., Lambert, S., Lecoq, J., Ledochowitsch, P., Lee, J. H., Leon, A., Li, Y., Liang, E., Long, F., Mace, K., Melchior, J., Millman, D., Mollenkopf, T., Nayan, C., Ng, L., Ngo, K., Nguyen, T., Nicovich, P. R., North, K., Ocker, G. K., Ollerenshaw, D., Oliver, M., Pachitariu, M., Perkins, J., Reding, M., Reid, D., Robertson, M., Ronellenfitch, K., Seid, S., Slaughterbeck, C., Stoecklin, M., Sullivan, D., Sutton, B., Swapp, J., Thompson, C., Turner, K., Wakeman, W., Whitesell, J. D., Williams, D., Williford, A., Young, R., Zeng, H., Naylor, S., Phillips, J. W., Reid, R. C., Mihalas, S., Olsen, S. R., and Koch, C. (2019). A survey of spiking activity reveals a functional hierarchy

of mouse corticothalamic visual areas. *bioRxiv*. 1

[57] Smith, M. A. and Kohn, A. (2008). Spatial and temporal scales of neuronal correlation in primary visual cortex. *Journal of Neuroscience*, 28(48):12591–12603. 2.4.2

[58] Ventura, V. (2004). Testing for and estimating latency effects for poisson and non-poisson spike trains. *Neural Computation*, 16(11):2323–2349. 2.2.3

[59] Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. A.2.1, A.2.1, A.2.1

[60] Vinci, G., Ventura, V., Smith, M. A., and Kass, R. E. (2018). Adjusted regularization of cortical covariance. *Journal of Computational Neuroscience*, 45(2):83–101. 1.3, 2.4.2, 3.3

[61] Xia, Y. and Li, L. (2017). Hypothesis testing of matrix graph model with application to brain connectivity analysis. *Biometrics*, 73(3):780–791. 1.4, 2

[62] Yang, Y., Aminoff, E., Tarr, M., and Robert, K. E. (2016). A state-space model of cross-region dynamic connectivity in meg/eeg. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 1234–1242. Curran Associates, Inc. 1, 1.5

[63] Ye, Y., Xia, Y., and Li, L. (2019). Paired Test of Matrix Graphs and Brain Connectivity Analysis. *arXiv e-prints*, page arXiv:1908.08095. 1.4

[64] Yin, J. and Li, H. (2012). Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, 107:119 – 140. 1.4

[65] Yu, B. M., Cunningham, J. P., Santhanam, G., Ryu, S. I., Shenoy, K. V., and Sahani, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635. PMID: 19357332. 1.5, 3.1, 3.1

[66] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67. 1.3, 2.2.1

[67] Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *The Annals of Statistics*, 42(2):532–562. 1.4, 3, B.1, B.1

[68] Zhu, Y. and Li, L. (2018). Multiple matrix gaussian graphs estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):927–950. 1.3, 1.4, 2.4.1, 2.4.1, 2.4.1

Appendix A

Appendix to Chapter 2

A.1 Additional Figures

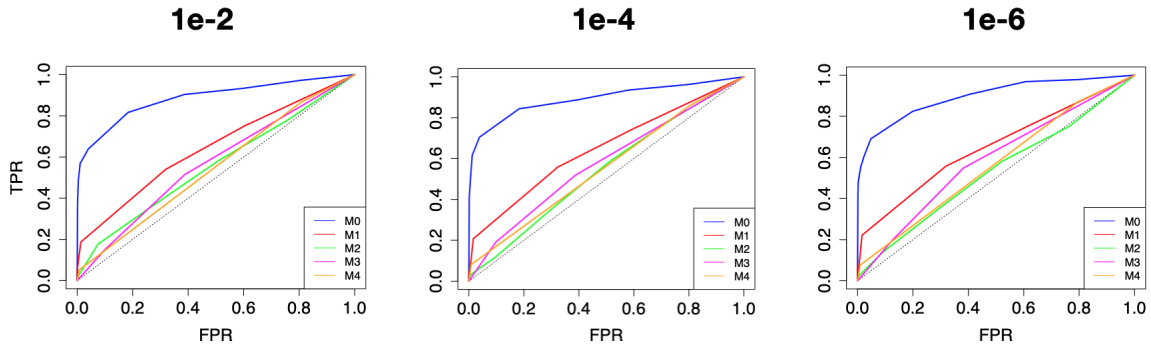

Figure A.1: ROC curve under random graph with $n = 5$, $p = 100$, $q = 30$ and $d = 5$. Each column corresponds to different tuning parameter value for our method (M0). Our method is consistently better regardless of tuning parameters.

A.2 Proofs

A.2.1 Technical Details

Here we introduced several lemmas that will help us prove propositions and theorems later.

Lemma 1. *For any $1 \leq i \leq j \leq q$, the sample estimate for (i, j) entry in column covariance of the t-th graph $\widehat{v}_{tij}^{samp} = \frac{\boldsymbol{Z}_{t,\cdot,i}^{\prime}\boldsymbol{Z}_{t,\cdot,j}}{n_t p}$ can be written as*

$$\widehat{v}_{tij}^{samp} = \frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\lambda_{tl}W_{tli}^{(k)}W_{tlj}^{(k)},$$

where λ_{tl} is the lth eigenvalue of $\boldsymbol{U}_t$, $(W_{tli}^{(k)}, W_{tlj}^{(k)}) \sim N(0, \boldsymbol{V}_{t,[i,j]})$ and are independent for all $1 \leq l \leq p, 1 \leq k \leq n_t$.

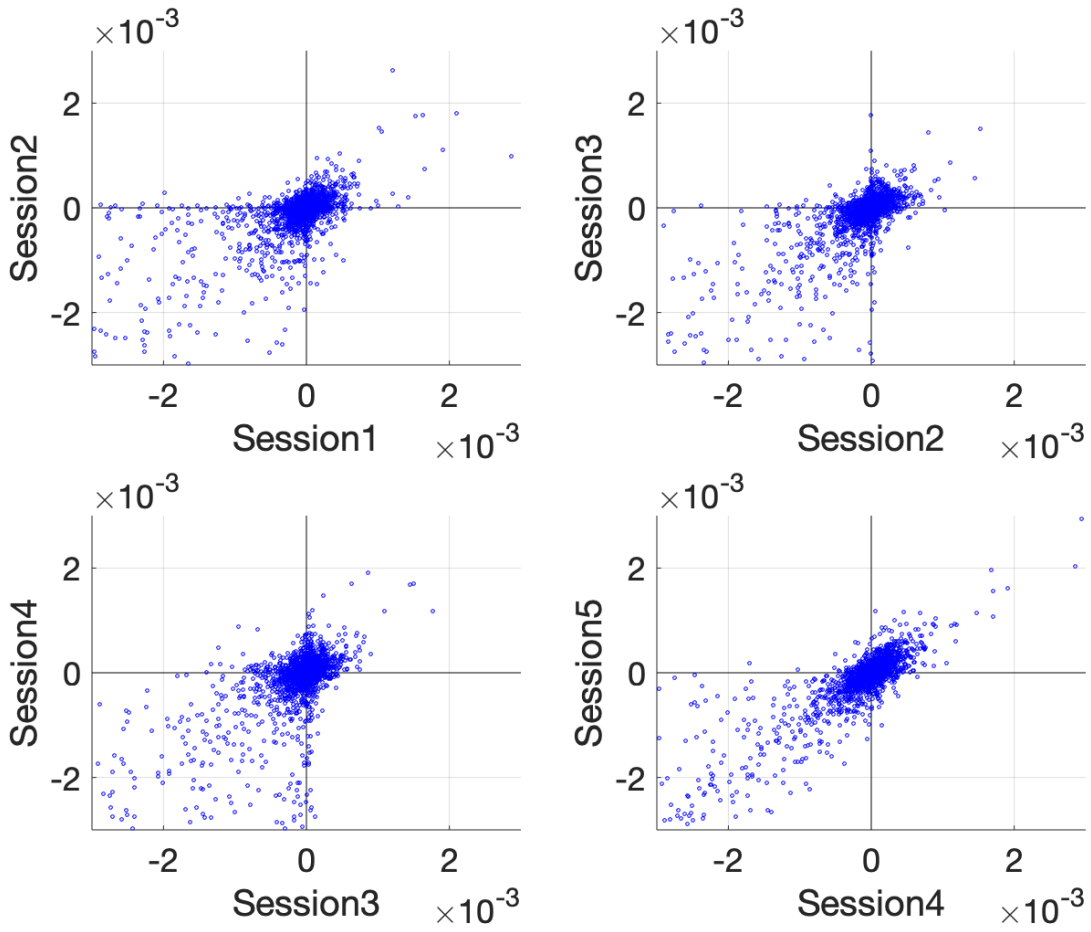

Figure A.2: Sample estimate of $\boldsymbol{B}_t$ for each session plotting against each other. For each panel, we plot vec($\widehat{\boldsymbol{B}}_t^{samp}$) vs. vec($\widehat{\boldsymbol{B}}_{t+1}^{samp}$), for $1 \leq t \leq 4$. Each blue dot corresponds to an edge value in precision matrix. The observation is similar to Figure 2.5 and the signs of connectivity should be consistent across sessions.

Proof. Let $\boldsymbol{Y}_t^{(k)} = \boldsymbol{U}_t^{-1/2}\boldsymbol{X}_t^{(k)} \in \mathbb{R}^p$, it immediately follows that $\boldsymbol{Y}_t^{(k)} \sim N(0, \boldsymbol{I}_{p\times p} \otimes \boldsymbol{V}_t)$, thus in particular,

$$(\boldsymbol{Y}_{t,\cdot,i}^{(k)}, \boldsymbol{Y}_{t,\cdot,j}^{(k)}) \sim N(0, \boldsymbol{I}_{p\times p} \otimes \boldsymbol{V}_{t,[i,j]}).$$

Assuming that we can decompose $\boldsymbol{U}_t = \boldsymbol{P}_t^{'}\boldsymbol{D}_t\boldsymbol{P}_t$, where $\boldsymbol{P}_t$ is an orthogonal matrix and $\boldsymbol{D}_t$ is a diagonal matrix with eigenvalues of $\boldsymbol{U}_t$ as its diagonal elements, we can define $\boldsymbol{W}_{t,\cdot,i}^{(k)} = \boldsymbol{P}_t\boldsymbol{Y}_{t,\cdot,i}^{(k)}$. Since $\boldsymbol{P}_t$ is orthogonal, we also have $(\boldsymbol{W}_{t,\cdot,i}^{(k)}, \boldsymbol{W}_{t,\cdot,j}^{(k)}) \sim N(0, \boldsymbol{I}_{p\times p} \otimes \boldsymbol{V}_{t,[i,j]})$, and $(W_{tli}^{(k)}, W_{tlj}^{(k)})$

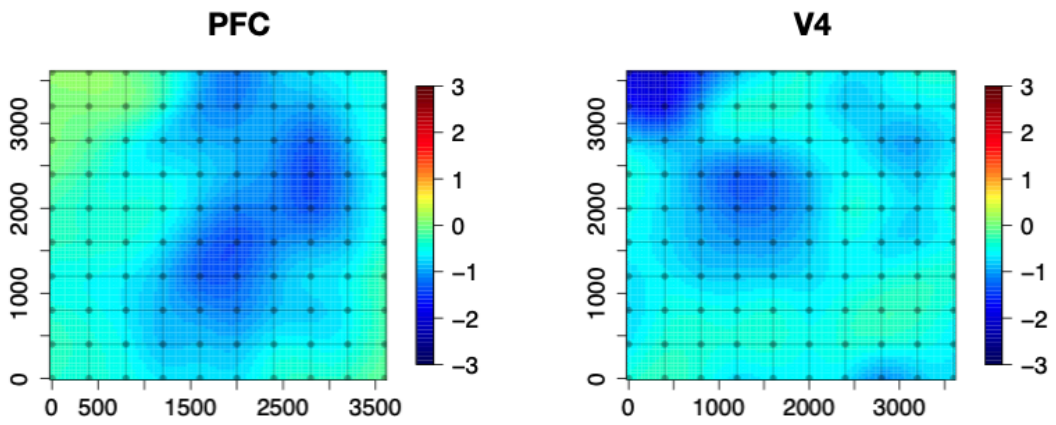

Figure A.3: Change of connectivity strength distribution over 2D array for PFC and V4 between late delay stage and cue at stage. X and Y axis are in micrometers. Black dots are the locations of electrodes in 2D array. Kernel smoother with 400 micrometer bandwidth is implemented to smooth out the signal with a resolution of 40×40 micrometer pixel. Values for 4 missing nodes on the array are interpolated with a Nadaraya-Watson normalization of the kernel. We can observe that most of the spatial changes are negative, meaning connectivity within region decreases during delay stage. Comparing V4 with PFC, we observe a dark blue region for V4 on the top left, which indicates that the negative change is larger for V4.

are independent $\forall 1 \leq l \leq p$. Thus we have

$$\begin{aligned}
\widehat{v}_{tij}^{samp} &= \frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p} X_{tli}^{(k)} X_{tlj}^{(k)} = \frac{1}{n_t p}\sum_{k=1}^{n_t} (\boldsymbol{U}_t^{-1/2}\boldsymbol{X}_{t,\cdot,i}^{(k)})' \boldsymbol{U}_t (\boldsymbol{U}_t^{-1/2}\boldsymbol{X}_{t,\cdot,j}^{(k)}) \\
&= \frac{1}{n_t p}\sum_{k=1}^{n_t} (\boldsymbol{P}_t\boldsymbol{Y}_{t,\cdot,i}^{(k)})' \boldsymbol{D}_t (\boldsymbol{P}_t\boldsymbol{Y}_{t,\cdot,j}^{(k)}) = \frac{1}{n_t p}\sum_{k=1}^{n_t} \boldsymbol{W}_{t,\cdot,i}^{(k)'} \boldsymbol{D}_t \boldsymbol{W}_{t,\cdot,j}^{(k)} \\
&= \frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p} \lambda_{tl} W_{tli}^{(k)} W_{tlj}^{(k)}.
\end{aligned}$$

□

Lemma 2. *For any $\delta > 0$, there exists a constant C which depends on c_e and δ only, such that*

$$\mathbb{P}\Big(\max_{1\leq t\leq d, 1\leq i\leq j\leq q} |\widehat{v}_{tij}^{samp} - V_{tij}| \geq C\sqrt{\frac{\log dq}{n_t p}}\Big) \leq (dq)^{-\delta}.$$

Proof. From Lemma 1, it's easy to see that $W_{tli}^{(k)} W_{tlj}^{(k)}$ is sub-exponential variable. From Assumption 2, Assumption 3 and Theorem 2.8.2 in [59], we prove our main claim. □

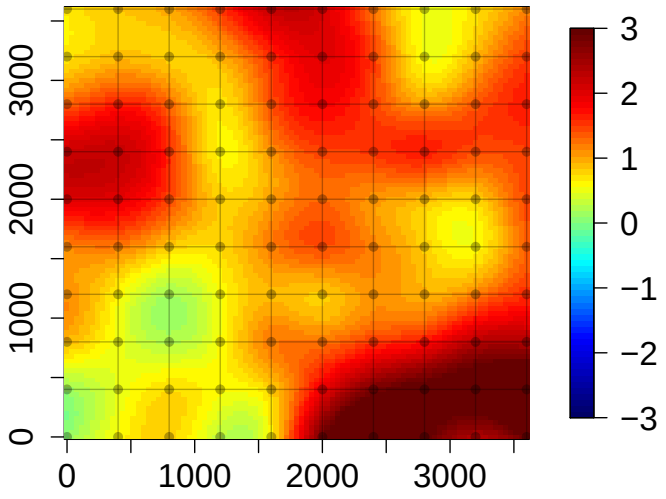

Figure A.4: Change of connectivity strength distribution over 2D array for V4 during cue stage between cue at $225°$ *and the rests. X and Y axis are in micrometers. Black dots are the locations of electrodes in 2D array. Kernel smoother with 400 micrometer bandwidth is implemented to smooth out the signal with a resolution of 40×40 micrometer pixel. Values for 4 missing nodes on the array are interpolated with a Nadaraya-Watson normalization of the kernel. We can observe that most of the spatial changes are positive, meaning a net increase of connectivity for cue at at* $225°$ *than the rests.*

Lemma 3. *Let* $\lambda_i = c\frac{\xi+1}{\xi-1}\sqrt{\frac{d+\log q}{n_0 p}}$ *for constant* $\xi > 1$ *and a large enough constant* c *depending on* $\delta,\ c_e$ *and* M_0 *only. Then with probability* $1 - q^{-\delta}$ *that event* E_i *holds where*

$$E_i = \{\max_{l\neq i} \frac{\left[\sum_{t=1}^{d}(\overline{\boldsymbol{Z}}'_{t,\cdot,l}\boldsymbol{E}_{ti})^2\right]^{1/2}}{n_0 p} \leq \frac{\xi-1}{\xi+1}\lambda_i\}.$$

Moreover, define $\widetilde{\boldsymbol{Z}}_{t,\cdot,i}\boldsymbol{D}_{ti}^{-1/2} = \overline{\widetilde{\boldsymbol{Z}}}_{t,\cdot,i}$*; with the same probability that event* $\widetilde{E}_i$ *holds where*

$$\widetilde{E}_i = \{\max_{1\leq i\leq q} \frac{\left[\sum_{t=1}^{d}(\overline{\widetilde{\boldsymbol{Z}}}'_{t,\cdot,i}\boldsymbol{E}_{ti})^2\right]^{1/2}}{n_0 p} \leq \frac{\xi-1}{\xi+1}\lambda_i\}.$$

Proof. Note that

$$(\boldsymbol{X}^{(k)}_{t,\cdot,l}, \boldsymbol{\varepsilon}^{(k)}_{t,\cdot,i}) \overset{iid}{\sim} N\left(\mathbf{0}, \boldsymbol{U}_t \otimes \text{diag}(v_{tll}, \frac{1}{b_{tii}})\right)$$

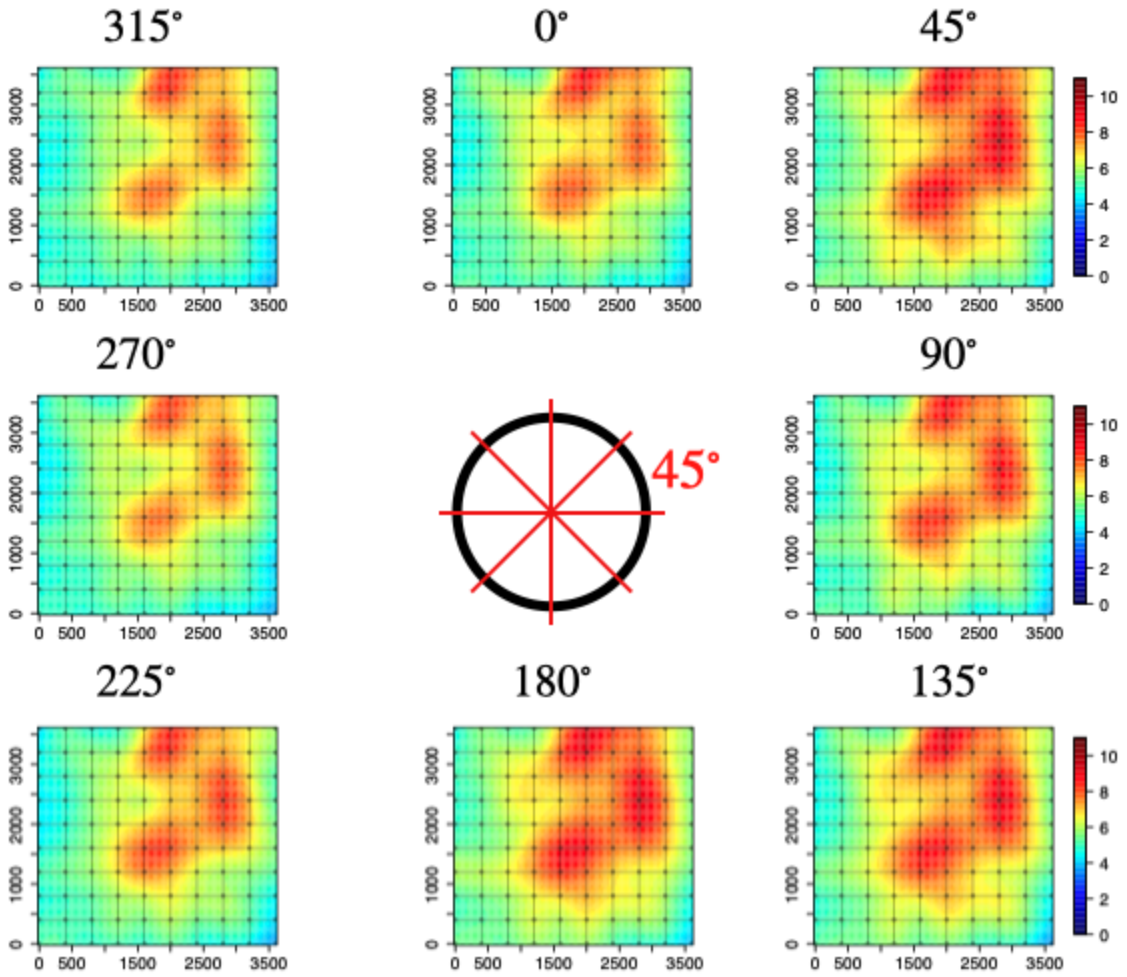

Figure A.5: Connectivity strength distribution over 2D array for PFC for eight different cues during cue stage. Eight different panel corresponds to eight different cues appearing at different angles. Notice that $45°/135°/180°$ *all show stronger connectivity than* $225°$.

for $1 \leq k \leq n_t$. Hence

$$(\boldsymbol{U}_t^{-1/2}\boldsymbol{X}_{t,\cdot,l}^{(k)}, \boldsymbol{U}_t^{-1/2}\boldsymbol{\varepsilon}_{t,\cdot,i}^{(k)}) \overset{iid}{\sim} N\left(\mathbf{0}, \boldsymbol{I} \otimes \text{diag}(v_{tll}, \frac{1}{b_{tii}})\right).$$

Condition on $\boldsymbol{Z}_{t,\cdot,l}$, we obtain that

$$\begin{aligned}\overline{\boldsymbol{Z}}'_{t,\cdot,l}\boldsymbol{E}_{ti} = D_{ti,ll}^{-1/2}\boldsymbol{Z}'_{t,\cdot,l}\boldsymbol{E}_{ti} &= D_{ti,ll}^{-1/2}\sum_{k=1}^{n_t}(\boldsymbol{X}_{t,\cdot,l}^{(k)})'\boldsymbol{U}_t^{1/2}\boldsymbol{U}_t^{-1/2}\boldsymbol{\varepsilon}_{t,\cdot,i}^{(k)} \\ &\sim N(0, \frac{g_{til}}{b_{tii}}),\end{aligned}$$

where $g_{til} = (\frac{\sum_{k=1}^{n_t}\|\boldsymbol{U}_t^{1/2}\boldsymbol{X}_{t,\cdot,l}^{(k)}\|_2^2}{\sum_{k=1}^{n_t}\|\boldsymbol{X}_{t,\cdot,l}^{(k)}\|_2^2}) \cdot n_t p$. In addition, due to the independence of d graphs, condition

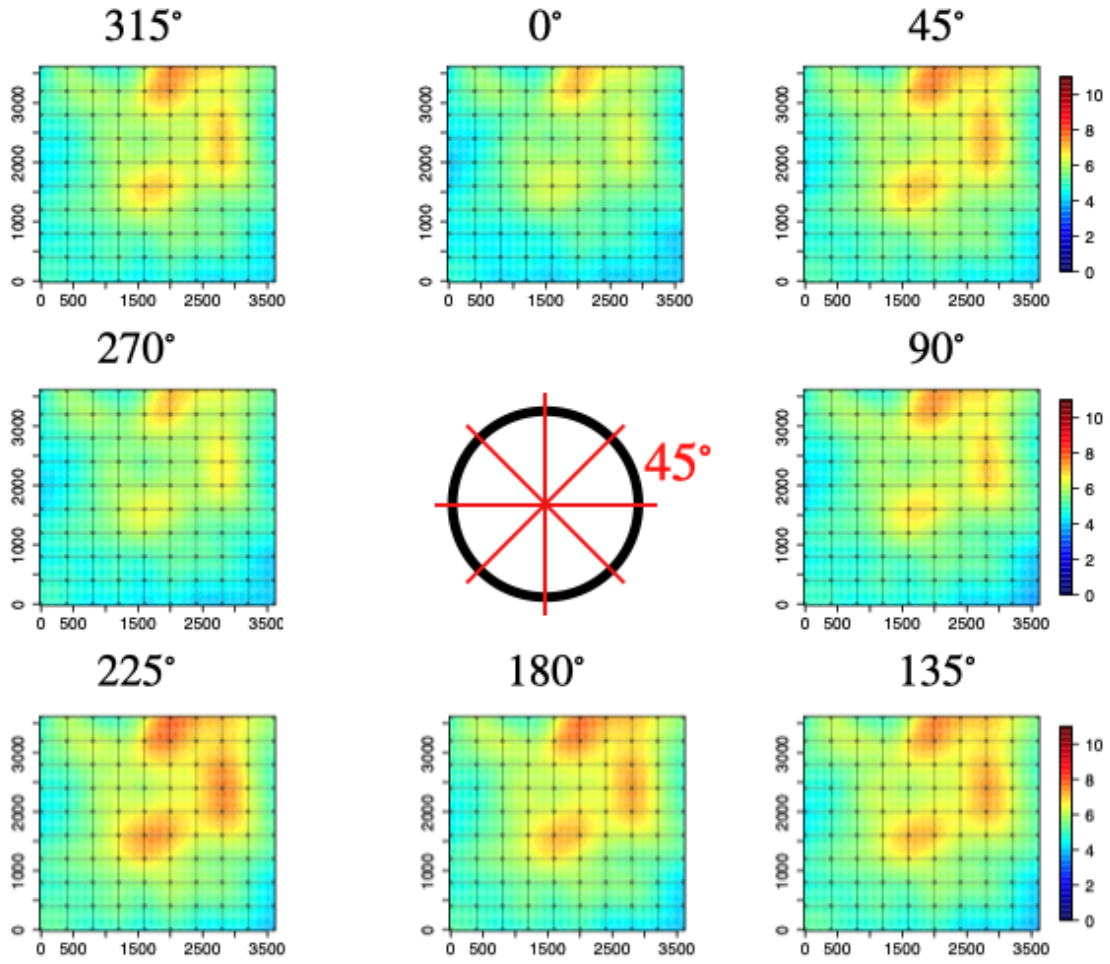

Figure A.6: Connectivity strength distribution over 2D array for PFC for eight different cues during late delay stage. Eight different panel corresponds to eight different cues appearing at different angles. The connectivity at $225°$ *is strongest among eight directions, which indicates that PFC may be impacted by V4.*

on $\{\boldsymbol{Z}_{t,\cdot,l}\}$, for $1 \le t \le d$, $\overline{\boldsymbol{Z}}'_{1,\cdot,l}\boldsymbol{E}_{1i}, \cdots, \overline{\boldsymbol{Z}}'_{d,\cdot,l}\boldsymbol{E}_{di}$ are independent. Consequently, $\overline{\boldsymbol{Z}}'_{t,\cdot,i}\boldsymbol{E}_{ti}\sqrt{\frac{b_{tii}}{g_{til}}}$ are i.i.d. $N(0,1)$, for $t \in d$ given $\{\boldsymbol{Z}_{t,\cdot,l}\}$ for $1 \le t \le d$. By applying the concentration inequality of χ^2 distribution (Lemma E.1 in [50] with $y = \delta_0 \log q$),

$$\mathbb{P}(\chi^2_d > d + 2\delta_0 \log q + 2\sqrt{\delta_0 d \log q}) \le q^{-\delta_0},$$

we obtain that

$$\mathbb{P}\left(\sum_{t=1}^{d}(\overline{\boldsymbol{Z}}'_{t,\cdot,l}\boldsymbol{E}_{ti})^2 > \max_t \frac{g_{til}}{b_{tii}} \cdot (d + 2\delta_0 \log q + 2\sqrt{\delta_0 d \log q}\right) \le q^{-\delta_0}. \tag{A.1}$$

Now we bound

$$\frac{g_{til}}{b_{tii}}\frac{1}{(n_0p)^2} \leq \frac{1}{n_0p}\frac{n_t}{n_0}\lambda_{tp}v_{tii}, \tag{A.2}$$

where λ_{tp} is the largest eigenvalue of $\boldsymbol{U}_t$. Finally we apply a union bound argument to Equation (A.1) over $1 \leq l \leq q$ with the help of Equation (A.2) to obtain that

$$\mathbb{P}\left(\max_{l\neq i}\frac{\left[\sum_{t=1}^{d}(\overline{\boldsymbol{Z}}'_{t,\cdot,i}\boldsymbol{E}_{ti})^2\right]^{1/2}}{n_0p} > \max_t(\frac{n_t\lambda_{tp}v_{tii}}{n_0})\sqrt{\frac{d+2\delta_0\log q+2\sqrt{\delta_0 d\log q}}{n_0p}}\right) \leq q^{-\delta_0+1}. \tag{A.3}$$

Let $\delta = \delta_0 - 1$, this immediately implies that E_i holds with probability at least $1 - q^{-\delta}$ with constant C depending on $\delta,\ c_e$ and M_0 only. In addition, the explicit formula of the bound can be useful for us to provide a data-driven procedure.

The second claim on $\widetilde{E}_i$ holds with the similar procedure by noticing that

$$(\boldsymbol{X}^{(k)}_{t,\cdot,-i}\boldsymbol{\beta}_{ti}, \boldsymbol{E}^{(k)}_{t,\cdot,i}) \overset{iid}{\sim} N\left(\boldsymbol{0}, \boldsymbol{U}_t \otimes \text{diag}(\frac{b_{tii}v_{tii}-1}{b_{tii}}, \frac{1}{b_{tii}})\right)$$

for $1 \leq k \leq n_t$. □

Lemma 4. *Define the event*

$$E_{dia} = \{|\frac{\boldsymbol{Z}'_{t,\cdot,i}\boldsymbol{Z}_{t,\cdot,i}}{n_tp}| \in (\frac{1}{2c_e}, 2c_e) \textit{ for all } 1 \leq t \leq d, 1 \leq i \leq q\},$$

then E_{dia} holds with probability $1 - q^{-\delta}$, where δ is a positive constant.

Proof. By Lemma 1 with $i = j$, we have

$$\frac{\boldsymbol{Z}'_{t,\cdot,i}\boldsymbol{Z}_{t,\cdot,i}}{n_tp} = \frac{1}{n_tp}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\lambda_{tl}(W^{(k)}_{tli})^2,$$

$W^{(k)}_{tli} \sim N(0, V_{tii})$ and are independent for all $1 \leq l \leq p, 1 \leq k \leq n_t$. Hence $\mathbb{E}\frac{\boldsymbol{Z}'_{t,\cdot,i}\boldsymbol{Z}_{t,\cdot,i}}{n_tp} = V_{tii}\frac{\boldsymbol{U}_t}{p} = V_{tii}$ by Assumption 2. By Assumption 3 and sub-exponential concentration inequality (e.g. Theorem 2.8.2 in [59]), combining with union bound and Assumption 1, we obtain that

$$\mathbb{P}\left(\max_{1\leq t\leq d, 1\leq i\in q}|\frac{\boldsymbol{Z}'_{t,\cdot,i}\boldsymbol{Z}_{t,\cdot,i}}{n_tp} - V_{tii}| \geq C\sqrt{\frac{\log dq}{n_0p}}\right) \leq (dq)^{-\delta},$$

where C depends on M_0, δ and c_e only. The desired result follows by noting that $V_{tii} \in (\frac{1}{c_e}, c_e)$ and $C\sqrt{\frac{\log dq}{n_0p}}$ is sufficiently small from Assumption 3 and 4 respectively. □

Lemma 5. *Define the event*

$$E_{com} = \{|\max_{l_1,l_2}\|\boldsymbol{M}_{l_1l_2}\|_2 \leq C\sqrt{\frac{\log(dq)}{n_0p}}\},$$

where $\boldsymbol{M}_{l_1,l_2} \in \mathbb{R}^{d\times d}$ *is a diagonal matrix with the* t*-th entry on the diagnal* $M_{l_1l_2,tt} = \frac{\boldsymbol{Z}'_{t,\cdot,l_1}\boldsymbol{Z}_{t,\cdot,l_2}}{n_0p} - \mathbb{E}\frac{\boldsymbol{Z}'_{t,\cdot,l_1}\boldsymbol{Z}_{t,\cdot,l_2}}{n_0p}$*, then* E_{com} *holds with probability* $1-(dq)^{-\delta}$*, where* δ *is a positive constant and* C *only depends on* M_0*,* δ *and* c_e*.*

Proof. Since $\boldsymbol{M}_{l_1l_2}$ is a diagnal matrix,

$$\max_{l_1,l_2}\|\boldsymbol{M}_{l_1l_2}\|_2 = \max_{l_1,l_2,t}|M_{l_1l_2,tt}| = \max_{l_1,l_2,t}|\frac{\boldsymbol{Z}'_{t,\cdot,l_1}\boldsymbol{Z}_{t,\cdot,l_2}}{n_0p} - \mathbb{E}\frac{\boldsymbol{Z}'_{t,\cdot,l_1}\boldsymbol{Z}_{t,\cdot,l_2}}{n_0p}|.$$

By Lemma 1, $\frac{\boldsymbol{Z}'_{t,\cdot,l_1}\boldsymbol{Z}_{t,\cdot,l_2}}{n_0p} = \frac{n_t}{n_0}\frac{1}{n_tp}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\lambda_{tl}W_{tli}^{(k)}W_{tlj}^{(k)}$, with $(W_{tli}^{(k)}, W_{tlj}^{(k)}) \sim N(0, \boldsymbol{V}_{t,[i,j]})$ and are independent for all $1\leq l\leq p, 1\leq k\leq n_t$. By Assumption 1 and 3, and sub-exponential concentration inequality (Theorem 2.8.2 in [59]), we have

$$\max_{l_1,l_2}\|\boldsymbol{M}_{l_1l_2}\|_2 \leq C\sqrt{\frac{\log(qd)}{n_tp}}\frac{n_t}{n_0} \leq C'\sqrt{\frac{\log(qd)}{n_0p}}$$

where the last inequality is due to $n_t \leq M_0n_0$ by Assumption 1. Note that constant C' only depends on δ, M_0 and c_e. □

Lemma 6. *For any* $\delta > 0$*, there exists a constant* C *which depends on* δ*,* c_e *only, such that*

$$\mathbb{P}(\max_{1\leq i\leq q}\max_{1\leq h\leq q, h\neq i}|\frac{1}{n_tp}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon_{tlj}^{(k)}X_{tlh}^{(k)}| \geq C\sqrt{\frac{\log q}{n_tp}}) \leq q^{-\delta},$$

and

$$\mathbb{P}(\max_{1\leq i\leq q}|\frac{1}{n_tp}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon_{tli}^{(k)}\boldsymbol{X}_{tl,-i}^{(k)}\boldsymbol{\beta}_{ti}| \geq C\sqrt{\frac{\log q}{n_tp}}) \leq q^{-\delta}.$$

Proof. Let $\boldsymbol{\varepsilon}_{ti}^{(k)} = (\varepsilon_{t1i}^{(k)}, \cdots, \varepsilon_{tpi}^{(k)})'$; with $h\neq i$, it's obvious that $\boldsymbol{\varepsilon}_{ti}^{(k)}$ is orthogonal to $\boldsymbol{X}_{t,\cdot,h}^{(k)}$. Therefore

$$\mathrm{Cov}((\boldsymbol{\varepsilon}_{ti}^{(k)}, \boldsymbol{X}_{t,\cdot,h}^{(k)})) = \boldsymbol{U}_t \otimes \mathrm{diag}((b_{tii})^{-1}, v_{thh}).$$

Following similar arguments in Lemma 2, we prove the first claim.

Let $\widetilde{\boldsymbol{X}}_{t,\cdot,i}^{(k)} = (\boldsymbol{X}_{t,1,-i}^{(k)}\boldsymbol{\beta}_{ti}, \cdots, \boldsymbol{X}_{t,p,-i}^{(k)}\boldsymbol{\beta}_{ti})'$, and $\boldsymbol{\varepsilon}_{ti}^{(k)}$ is orthogonal to $\widetilde{\boldsymbol{X}}_{ti}^{(k)}$. Notice that $\boldsymbol{X}_{t,\cdot,-i}^{(k)} \sim N(\boldsymbol{0}, \boldsymbol{U}_t \otimes \boldsymbol{V}_{t,-i,-i})$ and $\boldsymbol{\beta}_{ti} = -\frac{1}{b_{tii}}\boldsymbol{B}_{t,-i,i}$, therefore we have

$$\mathrm{Cov}(\widetilde{\boldsymbol{X}}_{t,\cdot,i}^{(k)}) = \frac{1}{b_{tii}^2}\mathrm{tr}(\boldsymbol{B}_{t,-i,i}\boldsymbol{B}_{t,i,-i}\boldsymbol{V}_{t,-i,-i})\boldsymbol{U}_t = \frac{b_{tii}v_{tii}-1}{b_{tii}}\boldsymbol{U}_t.$$

Then we have

$$\text{Cov}((\boldsymbol{\varepsilon}_{ti}^{(k)}, \widetilde{\boldsymbol{X}}_{t,\cdot,i}^{(k)})) = \boldsymbol{U}_t \otimes \text{diag}((b_{tii})^{-1}, \frac{b_{tii}v_{tii} - 1}{b_{tii}}).$$

Following similar arguments in Lemma 2, we prove the second claim. □

Lemma 7. *Define*

$$\begin{aligned}\widehat{\delta}_{tij} = &\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\widehat{\varepsilon}_{tli}^{(k)}\widehat{\varepsilon}_{tlj}^{(k)} - \frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon_{tli}^{(k)}\varepsilon_{tlj}^{(k)} \\ &- \frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\beta_{ti,j} - \widehat{\beta}_{ti,j})(\varepsilon_{tlj}^{(k)})^2\mathbb{1}(i \neq j) - \frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\beta_{tj,i} - \widehat{\beta}_{tj,i})(\varepsilon_{tli}^{(k)})^2\mathbb{1}(i \neq j).\end{aligned}$$

We conclude that

$$\sum_{t=1}^{d}|\widehat{\delta}_{tij}| = O(s\lambda_*^2).$$

Proof. Let $\boldsymbol{\gamma}_{ti} = (\beta_{ti,1}, ..., \beta_{ti,i-1}, -1, \beta_{ti,i+1}, ..., , \beta_{ti,q})'$; we have $\varepsilon_{tli}^{(k)} = -\boldsymbol{X}_{tl,\cdot}^{(k)}\boldsymbol{\gamma}_i$ and $\widehat{\varepsilon}_{tli}^{(k)} = -\boldsymbol{X}_{tl,\cdot}^{(k)}\widehat{\boldsymbol{\gamma}}_{ti}$. It follows that

$$\begin{aligned}&\sum_{t=1}^{d}\left(\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\widehat{\varepsilon}_{tli}^{(k)}\widehat{\varepsilon}_{tlj}^{(k)} - \frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon_{tli}^{(k)}\varepsilon_{tlj}^{(k)}\right) \\ &\qquad= \sum_{t=1}^{d}\left(\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\boldsymbol{\gamma}_{ti} - \widehat{\boldsymbol{\gamma}}_{ti})'(\boldsymbol{X}_{tl,\cdot}^{(k)})'\varepsilon_{tlj}^{(k)}\right. \qquad \text{(A.4)} \\ &\qquad+ \frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\boldsymbol{\gamma}_{tj} - \widehat{\boldsymbol{\gamma}}_{tj})'(\boldsymbol{X}_{tl,\cdot}^{(k)})'\varepsilon_{tli}^{(k)} \qquad \text{(A.5)} \\ &\qquad\left.+ \frac{1}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\boldsymbol{\gamma}_{ti} - \widehat{\boldsymbol{\gamma}}_{ti})'(\boldsymbol{X}_{tl,\cdot}^{(k)})'\boldsymbol{X}_{tl,\cdot}^{(k)}(\boldsymbol{\gamma}_{tj} - \widehat{\boldsymbol{\gamma}}_{tj})\right). \qquad \text{(A.6)}\end{aligned}$$

For the first and second term (Equation (A.4)- (A.5)), we have

$$
\begin{aligned}
&\sum_{t=1}^{d}\left(\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\boldsymbol{\gamma}_{ti}-\widehat{\boldsymbol{\gamma}}_{ti})'(\boldsymbol{X}_{tl,\cdot}^{(k)})'\varepsilon_{tlj}^{(k)}\right)\\
&\quad=\sum_{t=1}^{d}\left((\gamma_{ti,j}-\widehat{\gamma}_{ti,j})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}X_{tlj}^{(k)}\varepsilon_{tlj}^{(k)}\mathbb{1}(i\neq j)\right.\\
&\quad\left.+\sum_{h\neq i,h\neq j}(\gamma_{ti,h}-\widehat{\gamma}_{ti,h})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}X_{tlh}^{(k)}\varepsilon_{tlj}^{(k)}\right)\\
&\quad=\sum_{t=1}^{d}\left((\gamma_{ti,j}-\widehat{\gamma}_{ti,j})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\varepsilon_{tlj}^{(k)})^2\mathbb{1}(i\neq j)\right.\\
&\quad+(\gamma_{ti,j}-\widehat{\gamma}_{ti,j})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\boldsymbol{X}_{tl,-j}^{(k)}\boldsymbol{\beta}_{tj}\varepsilon_{tlj}^{(k)}\mathbb{1}(i\neq j)\\
&\quad\left.+\sum_{h\neq i,h\neq j}(\gamma_{ti,h}-\widehat{\gamma}_{ti,h})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}X_{tlh}^{(k)}\varepsilon_{tlj}^{(k)}\right)
\end{aligned}
$$

Thus

$$
\begin{aligned}
\sum_{t=1}^{d}\widehat{\delta}_{tij}=\sum_{t=1}^{d}\Bigg(&(\gamma_{ti,j}-\widehat{\gamma}_{ti,j})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\boldsymbol{X}_{tl,-j}^{(k)}\boldsymbol{\beta}_{tj}\varepsilon_{tlj}^{(k)}\mathbb{1}(i\neq j)\\
&+\sum_{h\neq i,h\neq j}(\gamma_{ti,h}-\widehat{\gamma}_{ti,h})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}X_{tlh}^{(k)}\varepsilon_{tlj}^{(k)}\\
&+(\gamma_{tj,i}-\widehat{\gamma}_{tj,i})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\boldsymbol{X}_{tl,-i}^{(k)}\boldsymbol{\beta}_{ti}\varepsilon_{tli}^{(k)}\mathbb{1}(i\neq j)\\
&+\sum_{h\neq i,h\neq j}(\gamma_{ti,h}-\widehat{\gamma}_{ti,h})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}X_{tlh}^{(k)}\varepsilon_{tlj}^{(k)}\\
&+\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\boldsymbol{\gamma}_{ti}-\widehat{\boldsymbol{\gamma}}_{ti})'(\boldsymbol{X}_{tl,\cdot}^{(k)})'\varepsilon_{tli}^{(k)}\\
&+\frac{1}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\boldsymbol{\gamma}_{ti}-\widehat{\boldsymbol{\gamma}}_{ti})'(\boldsymbol{X}_{tl,\cdot}^{(k)})'\boldsymbol{X}_{tl,\cdot}^{(k)}(\boldsymbol{\gamma}_{tj}-\widehat{\boldsymbol{\gamma}}_{tj})\Bigg)
\end{aligned}
\tag{A.7}
$$

Now noticing that

$$\sum_{t=1}^{d}\left|(\gamma_{ti,j}-\widehat{\gamma}_{ti,j})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\boldsymbol{X}_{tl,-j}^{(k)}\boldsymbol{\beta}_{tj}\varepsilon_{tlj}^{(k)}\mathbb{1}(i\neq j)\right| \leq \|\overline{\boldsymbol{\Delta}}_{i(j)}\|_2\cdot\max_{j\neq i}\frac{\left[\sum_{t=1}^{d}\left(\overline{\widetilde{\boldsymbol{Z}}}_{t,\cdot,j}^{'}\boldsymbol{\varepsilon}_{tj}\right)^2\right]^{1/2}}{n_t p}\mathbb{1}(i\neq j) \tag{A.8}$$

$$\sum_{t=1}^{d}\left|\sum_{h\neq i,h\neq j}(\gamma_{ti,h}-\widehat{\gamma}_{ti,h})\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}X_{tlh}^{(k)}\varepsilon_{tlj}^{(k)}\right| \leq \sum_{h\neq i,h\neq j}\|\overline{\boldsymbol{\Delta}}_{i(h)}\|_2\cdot\max_{h\neq i,h\neq j}\frac{[\sum_{t=1}^{d}(\overline{\boldsymbol{Z}}_{t,\cdot,h}^{'}\boldsymbol{\varepsilon}_{ti})^2]^{1/2}}{n_t p} \tag{A.9}$$

$$\begin{aligned}
&\sum_{t=1}^{d}\left|\left(\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\boldsymbol{\gamma}_{ti}-\widehat{\boldsymbol{\gamma}}_{ti})'(\boldsymbol{X}_{tl,\cdot}^{(k)})'\boldsymbol{X}_{tl,\cdot}^{(k)}(\boldsymbol{\gamma}_{tj}-\widehat{\boldsymbol{\gamma}}_{tj})\right)\right| \\
&\leq \max_{1\leq i\leq j\leq q}\sum_{t=1}^{d}|(\boldsymbol{\gamma}_{ti}-\widehat{\boldsymbol{\gamma}}_{ti})'\boldsymbol{V}_t(\boldsymbol{\gamma}_{tj}-\widehat{\boldsymbol{\gamma}}_{tj})| \\
&+\max_{1\leq i\leq j\leq q}\sum_{t=1}^{d}|(\boldsymbol{\gamma}_{ti}-\widehat{\boldsymbol{\gamma}}_{ti})'(\widehat{\boldsymbol{V}}_t-\boldsymbol{V}_t)(\boldsymbol{\gamma}_{tj}-\widehat{\boldsymbol{\gamma}}_{tj})| \\
&\leq c_e\max_{1\leq i\leq q}\|\boldsymbol{\beta}_i^0-\widehat{\boldsymbol{\beta}}_i^0\|_2^2+\max_{1\leq i\leq j\leq q}\|\boldsymbol{M}_{i,j}\|_2\cdot(\sum_{l\neq i}\|\boldsymbol{\beta}_{i(l)}-\widehat{\boldsymbol{\beta}}_{i(l)}\|_2)^2,
\end{aligned} \tag{A.10}$$

where the last inequality follows from that $\sum_{t=1}^{d}|(\boldsymbol{\gamma}_{ti}-\widehat{\boldsymbol{\gamma}}_{ti})'\boldsymbol{V}_t(\boldsymbol{\gamma}_{tj}-\widehat{\boldsymbol{\gamma}}_{tj})|\leq\max_{1\leq t\leq d}\lambda_{max}(\boldsymbol{V}_t)\|\boldsymbol{\beta}_i^0-\widehat{\boldsymbol{\beta}}_i^0\|_2^2$ and Assumption 3, and that the proof in norm compression inequality (Theorem 3.4 of [14]) and a simple fact that for any vector $\boldsymbol{v}\in\mathbb{R}^p$ and $\boldsymbol{M}\in\mathbb{R}^{p\times p}$, $\boldsymbol{v}'\boldsymbol{M}\boldsymbol{v}\leq\|\boldsymbol{v}\|_1^2|\boldsymbol{M}|_\infty$. Here $\boldsymbol{M}_{l_1,l_2}\in\mathbb{R}^{d\times d}$ and $\boldsymbol{M}_{l_1,l_2}=(m_{l_1,l_2,t_1,t_2})$, where $m_{l_1,l_2,t_1,t_2}=\frac{\boldsymbol{Z}_{t_1,\cdot,l_1}^{'}\boldsymbol{Z}_{t_2,\cdot,l_2}}{n_0 p}-\mathbb{E}\frac{\boldsymbol{Z}_{t_1,\cdot,l_1}^{'}\boldsymbol{Z}_{t_2,\cdot,l_2}}{n_0 p}$.

Combining Equation (A.7), (A.8), (A.9) and Equation (A.10), we finally have,

$$\begin{aligned}\sum_{t=1}^{d}|\widehat{\delta}_{tij}| \leq c_e \max_{1\leq i\leq q}\|\boldsymbol{\beta}_i^0-\widehat{\boldsymbol{\beta}}_i^0\|_2^2+\max_{1\leq i\leq j\leq q}\|\boldsymbol{M}_{i,j}\|_2\cdot(\sum_{l\neq i}\|\boldsymbol{\beta}_{i(l)}-\widehat{\boldsymbol{\beta}}_{i(l)}\|_2)^2\\ +\|\overline{\boldsymbol{\Delta}}_{i(j)}\|_2\cdot\max_{j\neq i}\frac{\left[\sum_{t=1}^{d}\left(\overline{\widetilde{\boldsymbol{Z}}}'_{t,\cdot,j}\boldsymbol{\varepsilon}_{tj}\right)^2\right]^{1/2}}{n_tp}\\ +\sum_{h\neq i,h\neq j}\|\overline{\boldsymbol{\Delta}}_{i(h)}\|_2\cdot\max_{h\neq i,h\neq j}\frac{[\sum_{t=1}^{d}(\overline{\boldsymbol{Z}}'_{t,\cdot,h}\boldsymbol{\varepsilon}_{ti})^2]^{1/2}}{n_tp}\\ +\|\overline{\boldsymbol{\Delta}}_{j(i)}\|_2\cdot\max_{i\neq j}\frac{\left[\sum_{t=1}^{d}\left(\overline{\widetilde{\boldsymbol{Z}}}'_{t,\cdot,i}\boldsymbol{\varepsilon}_{ti}\right)^2\right]^{1/2}}{n_tp}\\ +\sum_{h\neq i,h\neq j}\|\overline{\boldsymbol{\Delta}}_{i(h)}\|_2\cdot\max_{h\neq i,h\neq j}\frac{[\sum_{t=1}^{d}(\overline{\boldsymbol{Z}}'_{t,\cdot,h}\boldsymbol{\varepsilon}_{tj})^2]^{1/2}}{n_tp}. \qquad (A.11)\end{aligned}$$

We can bound the first term and second term with $O(s\lambda_*^2+s^2\lambda_*^3)=O(s\lambda_*^2)$ by Lemma 5 and Theorem 1, the rest terms with $s\lambda_*^2$ by Lemma 3 and Theorem 1. Therefore we finish our proof. □

Lemma 8. *Define* $\widetilde{r}_{tij}=\frac{1}{n_tp}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon_{tli}^{(k)}\varepsilon_{tlj}^{(k)}$*, then for any* $M>0$ *, there exists a constant C which depends on* δ*,* c_e*,* M_0 *only, such that*

$$\mathbb{P}\left(\max_{1\leq i\leq j\leq q}|\widetilde{r}_{tij}-\frac{b_{tij}}{b_{tii}b_{tjj}}|\geq C\sqrt{\frac{\log q}{n_tp}}\right)\leq q^{-\delta},$$

and

$$\mathbb{P}\left(\max_{1\leq i\leq j\leq q}|\sum_{t=1}^{d}(\widetilde{r}_{tij}-\frac{b_{tij}}{b_{tii}b_{tjj}})|\geq C\sqrt{\frac{d\log q}{n_0p}}\right)\leq q^{-\delta}.$$

Proof. From basic regression theory, it's immediate to see that $\mathrm{Cov}(\varepsilon_{tli}^{(k)})=\frac{u_{tll}}{b_{tii}}$ and $\mathrm{Cov}(\varepsilon_{tli}^{(k)},\varepsilon_{tlj}^{(k)})=u_{tll}\frac{b_{tij}}{b_{tii}b_{tjj}}$. By Lemma 1, we can write

$$\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon_{tli}^{(k)}\varepsilon_{tlj}^{(k)}=\sum_{k=1}^{n_t}\sum_{l=1}^{p}\lambda_{tl}\epsilon_{tli,}^{(k)}\epsilon_{tlj}^{(k)},$$

$$\sum_{t=1}^{d}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon_{tli}^{(k)}\varepsilon_{tlj}^{(k)}=\sum_{t=1}^{d}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\lambda_{tl}\epsilon_{tli}^{(k)}\epsilon_{tlj}^{(k)},$$

where $\epsilon_{tli}^{(k)}$ and $\epsilon_{tlj}^{(k)}$ are independent over k and l. Noticing that $\mathrm{E}(\varepsilon_{tli}^{(k)}\varepsilon_{tlj}^{(k)})=u_{tll}\frac{b_{tij}}{b_{tii}b_{tjj}}$, with Assumption 1 and following the same proof as in Lemma 2, we prove the two claims.

□

Lemma 9. *Assuming that $\boldsymbol{Y}_t \sim N(\boldsymbol{0}, \boldsymbol{U}_t)$, $\boldsymbol{X}_t \sim N(\boldsymbol{0}, \boldsymbol{U}_t \otimes \boldsymbol{V}_t)$, $\boldsymbol{A}_t = \boldsymbol{U}_t^{-1}$ and $\boldsymbol{B}_t = \boldsymbol{V}_t^{-1}$, with a predefined integer h, we define that*

$$\boldsymbol{c}_{ti}^h = \underset{\boldsymbol{\beta}_{ti}^h}{\operatorname{argmin}} \mathbb{E}(Y_{ti} - \boldsymbol{Y}_{t,i-h:i-1}^{'} \boldsymbol{\beta}_{ti}^h)^2,$$

$$d_{ti}^h = \mathit{Var}(\varepsilon_{ti}^h) = \mathit{Var}(Y_{ti} - \boldsymbol{Y}_{t,i-h:i-1}^{'} \boldsymbol{c}_{ti}^h),$$

$$\widetilde{\boldsymbol{c}}_{ti}^h = \underset{\boldsymbol{\beta}_{ti}^h}{\operatorname{argmin}} \frac{1}{q} \sum_{m=1}^{q} \mathbb{E}(X_{tim} - \boldsymbol{X}_{t,i-h:i-1,m}^{'} \boldsymbol{\beta}_{ti}^h)^2,$$

$$\widetilde{d}_{ti}^h = \frac{1}{q} \sum_{m=1}^{q} \mathit{Var}(\varepsilon_{tim}^h) = \frac{1}{q} \sum_{m=1}^{q} \mathit{Var}(X_{tim} - \boldsymbol{X}_{t,i-h:i-1,m}^{'} \widetilde{\boldsymbol{c}}_h^i),$$

we have

$$\widetilde{\boldsymbol{c}}_{ti}^h = \boldsymbol{c}_{ti}^h, \tag{A.12}$$

$$\widetilde{d}_{ti}^h = \frac{\mathit{Tr}(\boldsymbol{V})}{q} d_{ti}^h. \tag{A.13}$$

Proof. From regression theory, it immediately follows that

$$\boldsymbol{c}_{ti}^h = -(a_{h,h}^{t,h,i})^{-1} \boldsymbol{A}_{-h,h}^{t,h,i},$$

$$d_{ti}^h = (a_{h,h}^{t,h,i})^{-1},$$

$$\widetilde{\boldsymbol{c}}_{ti}^h = -(a_{h,h}^{t,h,i})^{-1} \boldsymbol{A}_{-h,h}^{t,h,i},$$

$$\widetilde{d}_{ti}^h = \frac{1}{q} \sum_{m=1}^{q} v_{tmm} (a_{h,h}^{t,h,i})^{-1} = \frac{\mathrm{Tr}(\boldsymbol{V}_t)}{q} (a_{h,h}^{t,h,i})^{-1}.$$

where $\boldsymbol{A}^{t,h,i} = \boldsymbol{U}_{t,i-h:i-1,i-h:i-1}^{-1} \in \mathbb{R}^{h \times h}$. Thus our claim holds. □

Lemma 10. *Following the procedure defined in Section 2.2.3, for any $\delta > 0$, there exists constants C which depends on δ, c_e only, such that*

$$\mathbb{P}\left(\max_{1 \le i \le p} \|\widehat{\boldsymbol{c}}_{ti} - \boldsymbol{c}_{ti}\|_2 \ge C \sqrt{\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}} \right) \le p^{-\delta},$$

$$\mathbb{P}\left(\max_{1 \le i \le p} |\widehat{d}_{ti} - \frac{\mathit{Tr}(\boldsymbol{V}_t)}{q} d_{ti}| \ge C \sqrt{\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}} \right) \le p^{-\delta}.$$

Proof. For clarity of notation, in this proof we use h instead of h_t, and recall that $h_t = \left\lceil (n_t q)^{\frac{1}{2\alpha_t+2}} \right\rceil$. Denote by

$$\widehat{\boldsymbol{\alpha}}_i = \frac{1}{n_t q} \sum_{k=1}^{n_t} \sum_{m=1}^{q} \boldsymbol{X}_{t,i-h:i-1,m}^{(k)} X_{tim}^{(k)},$$

From the fact that

$$\widehat{\boldsymbol{c}}_{ti}^{h} = \frac{1}{n_t q}(\widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1})^{-1} \sum_{k=1}^{n_t} \sum_{m=1}^{q} \boldsymbol{X}_{t,i-h:i-1,m}^{(k)} X_{tim}^{(k)},$$

it immediately follows that

$$\widehat{\boldsymbol{\alpha}}_i = \widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1} \widehat{\boldsymbol{c}}_{ti}^{h}. \tag{A.14}$$

Recall the sample estimate for $\widehat{d}_{ti}$ is defined as

$$\widehat{d}_{ti} = \frac{1}{n_t q} \sum_{k=1}^{n_t} \sum_{m=1}^{q} (\widehat{\varepsilon}_{tim}^{(k)})^2,$$

where $\widehat{\varepsilon}_{tim}^{(k)} = X_{tim}^{(k)} - \boldsymbol{X}_{t,i-h:i-1,m}^{(k)} \widehat{\boldsymbol{c}}_{ti}^{h}$. Denote by $\varepsilon_{tim}^{(k)} = X_{tim}^{(k)} - (\boldsymbol{X}_{t,i-h:i-1,m}^{(k)}) \widetilde{\boldsymbol{c}}_{ti}^{h}$; following Lemma 6, we have for any $\delta > 0$, there exists a constant C such that

$$\mathbb{P}\left(\max_{1 \le i \le p} \| \frac{1}{n_t q} \sum_{k=1}^{n_t} \sum_{m=1}^{q} \varepsilon_{tim}^{(k)} \boldsymbol{X}_{t,i-h:i-1,m}^{(k)} \|_\infty \ge C \sqrt{\frac{\log p}{n_t q}} \right) \le p^{-\delta}.$$

Combining with Equation (A.14), we have for any $\delta > 0$, there exists a constant C such that

$$\mathbb{P}\left(\max_{1 \le i \le p} \| \widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1} (\widehat{\boldsymbol{c}}_{ti}^{h} - \widetilde{\boldsymbol{c}}_{ti}^{h}) \|_\infty \ge C \sqrt{\frac{\log p}{n_t q}} \right) \le p^{-\delta}. \tag{A.15}$$

Denote $\Delta \boldsymbol{c}_{ti} = \widehat{\boldsymbol{c}}_{ti}^{h} - \widetilde{\boldsymbol{c}}_{ti}^{h}$, we have

$$\begin{aligned} \| \Delta \boldsymbol{c}_{ti}^{'} (\widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1} - \boldsymbol{U}_{t,i-h:i-1,i-h:i-1}) \Delta \boldsymbol{c}_{ti} \| &\le \| \widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1} - \boldsymbol{U}_{t,i-h:i-1,i-h:i-1} \|_\infty \| \Delta \boldsymbol{c}_{ti} \|_1^2 \\ &\le C \sqrt{\frac{\log p}{n_t q}} \| \Delta \boldsymbol{c}_{ti} \|_1^2 \\ &\le hC \sqrt{\frac{\log p}{n_t q}} \| \Delta \boldsymbol{c}_{ti} \|_2^2, \end{aligned} \tag{A.16}$$

with probability $1 - p^{-\delta}$ by Lemma 2 and Cauchy–Schwarz inequality $\| \Delta \boldsymbol{c}_{ti} \|_1 \le \sqrt{h} \| \Delta \boldsymbol{c}_{ti} \|_2$. By Assumption 3,

$$\Delta \boldsymbol{c}_{ti}^{'} \boldsymbol{U}_{t,i-h:i-1,i-h:i-1} \Delta \boldsymbol{c}_{ti} \ge \lambda_{min}(\boldsymbol{U}_{t,i-h:i-1,i-h:i-1}) \| \Delta \boldsymbol{c}_{ti} \|_2^2 \ge \frac{1}{c_e} \| \Delta \boldsymbol{c}_{ti} \|_2^2.$$

Therefore we have

$$\begin{aligned} \Delta \boldsymbol{c}_{ti}^{'} \widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1} \Delta \boldsymbol{c}_{ti} &= \Delta \boldsymbol{c}_{ti}^{'} (\widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1} - \boldsymbol{U}_{t,i-h:i-1,i-h:i-1}) \Delta \boldsymbol{c}_{ti} \\ &\quad + \Delta \boldsymbol{c}_{ti}^{'} \boldsymbol{U}_{t,i-h:i-1,i-h:i-1} \Delta \boldsymbol{c}_{ti} \ge \frac{1}{2c_e} \| \Delta \boldsymbol{c}_{ti} \|_2^2, \end{aligned} \tag{A.17}$$

with the same probability by noticing that $hC\sqrt{\frac{\log p}{n_t q}}\|\Delta \boldsymbol{c}_{ti}\|_2^2 < \frac{1}{2c_e}\|\Delta \boldsymbol{c}_{ti}\|_2^2$. Thus we have

$$\begin{aligned}\|\Delta \boldsymbol{c}_{ti}\|_2^2 &\leq 2c_e \Delta \boldsymbol{c}_{ti}^{'} \widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1} \Delta \boldsymbol{c}_{ti} \leq 2c_e \|\Delta \boldsymbol{c}_{ti}^{'} \widehat{\boldsymbol{U}}_{t,i-h:i-1,i-h:i-1}\|_\infty \|\Delta \boldsymbol{c}_{ti}\|_1 \\ &\leq 2c_e C\sqrt{\frac{\log p}{n_t q}}\sqrt{h}\|\Delta \boldsymbol{c}_{ti}\|_2,\end{aligned}$$

with probability $1 - p^{-\delta}$, where the last inequality follows from Equation (A.15) and $\|\Delta \boldsymbol{c}_{ti}\|_1 \leq \sqrt{h}\|\Delta \boldsymbol{c}_{ti}\|_2$. Therefore, we have for any $\delta > 0$, there exists a constant C, such that

$$\mathbb{P}\left(\max_{1\leq i\leq p} \|\widehat{\boldsymbol{c}}_{ti}^h - \widetilde{\boldsymbol{c}}_{ti}^h\|_2 \geq C\sqrt{\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}}\right) \leq p^{-\delta}. \tag{A.18}$$

We can similarly bound $\|\Delta \boldsymbol{c}_{ti}\|_1$ as for any $\delta > 0$, there exists a constant C, such that

$$\mathbb{P}\big(\max_{1\leq i\leq p} \|\widehat{\boldsymbol{c}}_{ti}^h - \widetilde{\boldsymbol{c}}_{ti}^h\|_1 \geq C\sqrt{\frac{\log p,}{(n_t q)^{\frac{\alpha_t}{\alpha_t+1}}}}\big) \leq p^{-\delta}.$$

Following similar proof as in Lemma 8, we can show that for any $\delta > 0$, there exists constants C, such that

$$\mathbb{P}\left(\max_{1\leq i\leq p} |\frac{1}{n_t q}\sum_{k=1}^{n_t}\sum_{m=1}^{q}(\varepsilon_{tim}^{(k)})^2 - \widetilde{d}_{ti}^h| \geq C\sqrt{\frac{\log p}{n_t q}}\right) \leq p^{-\delta}. \tag{A.19}$$

Following similar proof as in Lemma 7, it's easy to verify that that for any $\delta > 0$, there exists constants C, such that

$$\mathbb{P}\left(\max_{1\leq i\leq p} |\frac{1}{n_t q}\sum_{k=1}^{n_t}\sum_{m=1}^{q}(\widehat{\varepsilon}_{tim}^{(k)})^2 - \frac{1}{n_t q}\sum_{k=1}^{n_t}\sum_{m=1}^{q}(\varepsilon_{tim}^{(k)})^2| \geq C\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}\right) \leq p^{-\delta}. \tag{A.20}$$

Combining the above two inequalities in Equation (A.19) and (A.20), and noticing that both $\sqrt{\frac{\log p}{n_t q}}$ and $\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}$ are asymptotically smaller than $\sqrt{\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}}$, we conclude that

$$\mathbb{P}\left(\max_{1\leq i\leq p} |\widehat{d}_{ti} - \widetilde{d}_{ti}^h| \geq C\sqrt{\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}}\right) \leq p^{-\delta}. \tag{A.21}$$

As in Lemma B.3 in [40],

$$\begin{aligned}|d_{ti}^h - d_{ti}|^2 &\leq C(n_t q)^{-\frac{2\alpha_t+1}{2\alpha_t+2}}, \\ \|\boldsymbol{c}_{ti}^* - \boldsymbol{c}_{ti}\|_2^2 &\leq C(n_t q)^{-\frac{2\alpha_t+1}{2\alpha_t+2}},\end{aligned}$$

where $\boldsymbol{c}_{ti}^*$ is a zero-padded vector of length $i-1$ with first $i-h-1$ elements being zero and later elements being $\boldsymbol{c}_{ti}^h$. With Equation (A.12), Equation (A.13), Equation (A.18) and Equation (A.21), we have for any $\delta > 0$, there exists constants C depending on δ, c_e only, such that

$$\mathbb{P}\left(\max_{1\leq i\leq p} \|\widehat{\boldsymbol{c}}_{ti} - \boldsymbol{c}_{ti}\|_2 \geq C\sqrt{\frac{\log p}{(n_t q)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}}\right) \leq p^{-\delta}, \tag{A.22}$$

$$\mathbb{P}\left(\max_{1\le i\le p}|\widehat{d}_{ti}-\frac{\mathrm{Tr}(\boldsymbol{V}_t)}{q}d_{ti}|\ge C\sqrt{\frac{\log p}{(n_tq)^{\frac{2\alpha_t+1}{2\alpha_t+2}}}}\right)\le p^{-\delta}. \tag{A.23}$$

□

Lemma 11. *For each element $\boldsymbol{W}^P$ as defined in Equation* (2.16)*, we have*

$$\begin{aligned}
w^P_{tij}=\frac{1}{d}\sum_{t=1}^{d}\big(&\frac{\|\boldsymbol{U}_t\|_F^2}{p}(\rho_{t,\chi_1(i),\chi_1(j)}\rho_{t,\chi_2(i),\chi_2(j)}+\rho_{t,\chi_1(i),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(i)}\\
&+\frac{1}{2}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(j),\chi_2(j)}\rho^2_{t,\chi_1(i),\chi_1(j)}+\frac{1}{2}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(j),\chi_2(j)}\rho^2_{t,\chi_2(i),\chi_2(j)}\\
&+\frac{1}{2}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(j),\chi_2(j)}\rho^2_{t,\chi_1(i),\chi_2(j)}+\frac{1}{2}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(j),\chi_2(j)}\rho^2_{t,\chi_1(j),\chi_2(i)}\\
&-\rho_{t,\chi_1(i),\chi_1(j)}\rho_{t,\chi_1(j),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(i)}-\rho_{t,\chi_1(i),\chi_1(j)}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(i),\chi_2(j)}\\
&-\rho_{t,\chi_2(i),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(j)}\rho_{t,\chi_1(i),\chi_2(j)}-\rho_{t,\chi_2(i),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(i)}\rho_{t,\chi_1(i),\chi_2(i)}\big)\big).
\end{aligned}$$

Proof. First, it's easy to observe that

$$\begin{aligned}
\boldsymbol{W}^P&=\mathbb{E}\{(\frac{1}{\sqrt{d}}\sum_{k=1}^{d}\sqrt{n_tp}\boldsymbol{\xi}_{t\mathcal{S}}\circ\boldsymbol{\Theta}_{tS})(\frac{1}{\sqrt{d}}\sum_{k=1}^{d}\boldsymbol{\xi}_{t\mathcal{S}}\circ\sqrt{n_tp}\boldsymbol{\Theta}_{tS})'\}\\
&=\frac{1}{d}\sum_{k=1}^{d}n_tp\boldsymbol{W}_t^{\rho}.
\end{aligned}\tag{A.24}$$

Therefore, we simply need to calculate each entry in $\boldsymbol{W}_t^{\rho}=\mathbb{E}\{\boldsymbol{\Theta}_{tS}(\boldsymbol{\Theta}_{tS})'\}$.

To proceed for the calculation of $\boldsymbol{W}_t^{\rho}$, we omit t here since the calculation is the same for each graph. Denote by $\widetilde{r}_{tij}=\frac{1}{n_tp}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon^{(k)}_{tli}\varepsilon^{(k)}_{tlj}$, and

$$\widetilde{\delta}_{tij}=\widetilde{r}_{tij}-r_{tij}=\frac{1}{n_tp}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon^{(k)}_{tli}\varepsilon^{(k)}_{tlj}-\frac{b_{tij}}{b_{tii}b_{tjj}}. \tag{A.25}$$

For each entry W^{ρ}_{ij}, we have

$$\begin{aligned}
npW^{\rho}_{ij}=\frac{1}{np}\mathbb{E}(&\sum_{k=1}^{n}\sum_{l=1}^{p}\frac{\varepsilon^{(k)}_{l\chi_1(i)}\varepsilon^{(k)}_{l\chi_2(i)}-r^l_{\boldsymbol{\chi(i)}}}{\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}}}-\frac{r_{\chi_1(i),\chi_2(i)}}{2r_{\chi_1(i),\chi_1(i)}}\frac{(\varepsilon^{(k)}_{l\chi_1(i)})^2-r^l_{\chi_1(i),\chi_1(i)}}{\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}}}\\
&-\frac{r_{\chi_1(i),\chi_2(i)}}{2r_{\chi_2(i),\chi_2(i)}}\frac{(\varepsilon^{(k)}_{l\chi_2(i)})^2-r^l_{\chi_2(i),\chi_2(i)}}{\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}}})(\sum_{k=1}^{n}\sum_{l=1}^{p}\frac{\varepsilon^{(k)}_{l\chi_1(j)}\varepsilon^{(k)}_{l\chi_2(j)}-r^l_{\boldsymbol{\chi(j)}}}{\sqrt{r_{\chi_1(j),\chi_1(j)}r_{\chi_2(j),\chi_2(j)}}}\\
&-\frac{r_{\chi_1(j),\chi_2(j)}}{2r_{\chi_1(j),\chi_1(j)}}\frac{(\varepsilon^{(k)}_{l\chi_1(j)})^2-r^l_{\chi_1(j),\chi_1(j)}}{\sqrt{r_{\chi_1(j),\chi_1(j)}r_{\chi_2(j),\chi_2(j)}}}-\frac{r_{\chi_1(j),\chi_2(j)}}{2r_{\chi_2(j),\chi_2(j)}}\frac{(\varepsilon^{(k)}_{l\chi_2(j)})^2-r^l_{\chi_2(j),\chi_2(j)}}{\sqrt{r_{\chi_1(j),\chi_1(j)}r_{\chi_2(j),\chi_2(j)}}})
\end{aligned}$$

Now applying Lemma 1,

$$\sum_{k=1}^{n}\sum_{l=1}^{p}(\varepsilon^{(k)}_{l\chi_1(i)}\varepsilon^{(k)}_{l\chi_2(i)}-r^l_{\boldsymbol{\chi(i)}})=\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l(\epsilon^{(k)}_{l\chi_1(i)}\epsilon^{(k)}_{l\chi_2(i)}-r_{\boldsymbol{\chi(i)}})$$

$$\sum_{k=1}^{n}\sum_{l=1}^{p}((\varepsilon_{l\chi_1(i)}^{(k)})^2 - r_{\chi_1(i),\chi_1(i)}^{l}) = \sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l((\epsilon_{l\chi_1(i)}^{(k)})^2 - r_{\chi_1(i),\chi_1(i)})$$

It's easy to show that

$$\begin{aligned}
npW_{ij}^{\rho} = &\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_2(i)}^{(k)}\epsilon_{l\chi_1(j)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)}) - r_{\boldsymbol{\chi_{(i)}}}r_{\boldsymbol{\chi_{(j)}}}}{\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\\
&+\frac{r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_1(j)}^{(k)})^2 - r_{\chi_1(i),\chi_1(i)}r_{\chi_1(j),\chi_1(j)})}{4r_{\chi_1(i),\chi_1(i)}r_{\chi_1(j),\chi_1(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\\
&+\frac{r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_2(i)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)})^2 - r_{\chi_2(i),\chi_2(i)}r_{\chi_2(j),\chi_2(j)})}{4r_{\chi_2(i),\chi_2(i)}r_{\chi_2(j),\chi_2(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\\
&-\frac{r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_2(i)}^{(k)}(\epsilon_{l\chi_1(j)}^{(k)})^2) - r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_1(j)})}{2r_{\chi_1(j),\chi_1(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\\
&-\frac{r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_1(j)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)}(\epsilon_{l\chi_1(i)}^{(k)})^2) - r_{\chi_1(j),\chi_2(j)}r_{\chi_1(i),\chi_1(i)})}{2r_{\chi_1(i),\chi_1(i)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\\
&-\frac{r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_2(i)}^{(k)}(\epsilon_{l\chi_2(j)}^{(k)})^2) - r_{\chi_1(i),\chi_2(i)}r_{\chi_2(j),\chi_2(j)})}{2r_{\chi_2(j),\chi_2(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\\
&-\frac{r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_1(j)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)}(\epsilon_{l\chi_2(i)}^{(k)})^2) - r_{\chi_1(i),\chi_2(i)}r_{\chi_2(i),\chi_2(i)})}{2r_{\chi_2(i),\chi_2(i)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\\
&+\frac{r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)})^2 - r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)})}{4r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\\
&+\frac{r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_2(i)}^{(k)}\epsilon_{l\chi_1(j)}^{(k)})^2 - r_{\chi_2(i),\chi_2(i)}r_{\chi_1(j),\chi_1(j)})}{4r_{\chi_2(i),\chi_2(i)}r_{\chi_1(j),\chi_1(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)\\
&= M_1 + M_2 + M_3 + M_4 + M_5 + M_6 + M_7 + M_8 + M_9,
\end{aligned}$$

where

$$\begin{aligned}
M_1 &= \frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_2(i)}^{(k)}\epsilon_{l\chi_1(j)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)}) - r_{\boldsymbol{\chi_{(i)}}}r_{\boldsymbol{\chi_{(j)}}}}{\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)\\
&= \frac{\|\boldsymbol{U}\|_F^2}{p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\big(b_{\chi_1(i),\chi_1(j)}b_{\chi_2(i),\chi_2(j)} + b_{\chi_1(i),\chi_2(j)}b_{\chi_1(j),\chi_2(i)}\big),
\end{aligned}$$

$$\begin{aligned}
M_2 &= \frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_2(j)}(\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_1(j)}^{(k)})^2 - r_{\chi_1(i),\chi_1(i)}r_{\chi_1(j),\chi_1(j)})}{4r_{\chi_1(i),\chi_1(i)}r_{\chi_1(j),\chi_1(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)\\
&= \frac{\|\boldsymbol{U}\|_F^2}{2p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\frac{b_{\chi_1(i),\chi_2(i)}b_{\chi_1(j),\chi_2(j)}b_{\chi_1(i),\chi_1(j)}^2}{b_{\chi_1(i),\chi_1(i)}b_{\chi_1(j),\chi_1(j)}},
\end{aligned}$$

$$M_3 = \frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_2(j)}\big(\mathbb{E}(\epsilon_{l\chi_2(i)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)})^2 - r_{\chi_2(i),\chi_2(i)}r_{\chi_2(j),\chi_2(j)}\big)}{4r_{\chi_2(i),\chi_2(i)}r_{\chi_2(j),\chi_2(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)$$
$$= \frac{\|\boldsymbol{U}\|_F^2}{2p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\frac{b_{\chi_1(i),\chi_2(i)}b_{\chi_1(j),\chi_2(j)}b_{\chi_2(i),\chi_2(j)}^2}{b_{\chi_2(i),\chi_2(i)}b_{\chi_2(j),\chi_2(j)}},$$

$$M_4 = -\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{r_{\chi_1(j),\chi_2(j)}\big(\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_2(i)}^{(k)}(\epsilon_{l\chi_1(j)}^{(k)})^2) - r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_1(j)}\big)}{2r_{\chi_1(j),\chi_1(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)$$
$$= -\frac{\|\boldsymbol{U}\|_F^2}{p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\frac{b_{\chi_1(j),\chi_2(j)}}{b_{\chi_1(j),\chi_1(j)}}b_{\chi_1(i),\chi_1(j)}b_{\chi_1(j),\chi_2(i)},$$

$$M_5 = -\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{r_{\chi_1(j),\chi_2(j)}\big(\mathbb{E}(\epsilon_{l\chi_1(j)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)}(\epsilon_{l\chi_1(i)}^{(k)})^2) - r_{\chi_1(j),\chi_2(j)}r_{\chi_1(i),\chi_1(i)}\big)}{2r_{\chi_1(i),\chi_1(i)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)$$
$$= -\frac{\|\boldsymbol{U}\|_F^2}{p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\frac{b_{\chi_1(i),\chi_2(i)}}{b_{\chi_1(i),\chi_1(i)}}b_{\chi_1(i),\chi_1(j)}b_{\chi_1(i),\chi_2(j)},$$

$$M_6 = \frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{r_{\chi_1(j),\chi_2(j)}\big(\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_2(i)}^{(k)}(\epsilon_{l\chi_2(j)}^{(k)})^2) - r_{\chi_1(i),\chi_2(i)}r_{\chi_2(j),\chi_2(j)}\big)}{2r_{\chi_2(j),\chi_2(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)$$
$$= -\frac{\|\boldsymbol{U}\|_F^2}{p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\frac{b_{\chi_1(j),\chi_2(j)}}{b_{\chi_2(j),\chi_2(j)}}b_{\chi_1(i),\chi_2(j)}b_{\chi_2(i),\chi_2(j)},$$

$$M_7 = -\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{r_{\chi_1(j),\chi_2(j)}\big(\mathbb{E}(\epsilon_{l\chi_1(j)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)}(\epsilon_{l\chi_2(i)}^{(k)})^2) - r_{\chi_1(i),\chi_2(i)}r_{\chi_2(i),\chi_2(i)}\big)}{2r_{\chi_2(i),\chi_2(i)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)$$
$$= -\frac{\|\boldsymbol{U}\|_F^2}{p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\frac{b_{\chi_1(i),\chi_2(i)}}{b_{\chi_2(i),\chi_2(i)}}b_{\chi_1(j),\chi_2(i)}b_{\chi_2(i),\chi_2(j)},$$

$$M_8 = \frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_2(j)}\big(\mathbb{E}(\epsilon_{l\chi_1(i)}^{(k)}\epsilon_{l\chi_2(j)}^{(k)})^2 - r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}\big)}{4r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)$$
$$= \frac{\|\boldsymbol{U}\|_F^2}{2p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\frac{b_{\chi_1(i),\chi_2(i)}b_{\chi_1(j),\chi_2(j)}b_{\chi_1(i),\chi_2(j)}^2}{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}},$$

$$M_9 = \frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\lambda_l^2\Big(\frac{r_{\chi_1(i),\chi_2(i)}r_{\chi_1(j),\chi_2(j)}\big(\mathbb{E}(\epsilon_{l\chi_2(i)}^{(k)}\epsilon_{l\chi_1(j)}^{(k)})^2 - r_{\chi_2(i),\chi_2(i)}r_{\chi_1(j),\chi_1(j)}\big)}{4r_{\chi_2(i),\chi_2(i)}r_{\chi_1(j),\chi_1(j)}\sqrt{r_{\chi_1(i),\chi_1(i)}r_{\chi_2(i),\chi_2(i)}r_{\chi_1(i),\chi_1(i)}r_{\chi_2(j),\chi_2(j)}}}\Big)$$
$$= \frac{\|\boldsymbol{U}\|_F^2}{2p\sqrt{b_{\chi_1(i),\chi_1(i)}b_{\chi_2(i),\chi_2(i)}b_{\chi_1(i),\chi_1(i)}b_{\chi_2(j),\chi_2(j)}}}\frac{b_{\chi_1(i),\chi_2(i)}b_{\chi_1(j),\chi_2(j)}b_{\chi_1(j),\chi_2(i)}^2}{b_{\chi_1(j),\chi_1(j)}b_{\chi_2(i),\chi_2(i)}}.$$

Combing all nine terms and noticing that $\rho_{i,j} = \frac{b_{i,j}}{\sqrt{b_{i,i}b_{j,j}}}$ we have

$$\begin{aligned}
npW_{ij}^{\rho} = \frac{\|\boldsymbol{U}\|_F^2}{p}(&\rho_{\chi_1(i),\chi_1(j)}\rho_{\chi_2(i),\chi_2(j)} + \rho_{\chi_1(i),\chi_2(j)}\rho_{\chi_1(j),\chi_2(i)} \\
&+ \frac{1}{2}\rho_{\chi_1(i),\chi_2(i)}\rho_{\chi_1(j),\chi_2(j)}\rho_{\chi_1(i),\chi_1(j)}^2 + \frac{1}{2}\rho_{\chi_1(i),\chi_2(i)}\rho_{\chi_1(j),\chi_2(j)}\rho_{\chi_2(i),\chi_2(j)}^2 \\
&+ \frac{1}{2}\rho_{\chi_1(i),\chi_2(i)}\rho_{\chi_1(j),\chi_2(j)}\rho_{\chi_1(i),\chi_2(j)}^2 + \frac{1}{2}\rho_{\chi_1(i),\chi_2(i)}\rho_{\chi_1(j),\chi_2(j)}\rho_{\chi_1(j),\chi_2(i)}^2 \\
&- \rho_{\chi_1(i),\chi_1(j)}\rho_{\chi_1(j),\chi_2(j)}\rho_{\chi_1(j),\chi_2(i)} - \rho_{\chi_1(i),\chi_1(j)}\rho_{\chi_1(i),\chi_2(i)}\rho_{\chi_1(i),\chi_2(j)} \\
&- \rho_{\chi_2(i),\chi_2(j)}\rho_{\chi_1(j),\chi_2(j)}\rho_{\chi_1(i),\chi_2(j)} - \rho_{\chi_2(i),\chi_2(j)}\rho_{\chi_1(j),\chi_2(i)}\rho_{\chi_1(i),\chi_2(i)}).
\end{aligned}$$

Then by Equation (A.24) we finish our proof. □

A.3 Proof of the Propositions

Proof of Proposition 1

Proof. First, recall that $r_{tij} = \frac{b_{tij}}{b_{tii}b_{tjj}}$, $\widetilde{\delta}_{tij} = \widetilde{r}_{tij} - r_{tij}$. Starting with $i = j$, we have

$$\widehat{\delta}_{tii} = \frac{1}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\widehat{\varepsilon}_{tli}^{(k)}\widehat{\varepsilon}_{tli}^{(k)} - \frac{1}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}\varepsilon_{tli}^{(k)}\varepsilon_{tli}^{(k)} = \widehat{r}_{tii} - \widetilde{r}_{tii}. \tag{A.26}$$

So we have

$$\widehat{r}_{tii} - r_{tii} = \widetilde{\delta}_{tii} + \widehat{\delta}_{tii}.$$

By Equation (2.7), denote by $\Delta_{\varepsilon,\beta}^{tij} = \frac{1}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\widehat{\beta}_{tij}((\widehat{\varepsilon}_{tlj}^{(k)})^2 - (\varepsilon_{tlj}^{(k)})^2) + \widehat{\beta}_{tji}((\widehat{\varepsilon}_{tli}^{(k)})^2 - (\varepsilon_{tli}^{(k)})^2))$, we have

$$\begin{aligned}
r_{tij} - \widehat{r}_{tij} &= r_{tij} + \frac{1}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\widehat{\varepsilon}_{tli}^{(k)}\widehat{\varepsilon}_{tlj}^{(k)} + \widehat{\beta}_{tij}(\widehat{\varepsilon}_{tlj}^{(k)})^2 + \widehat{\beta}_{tji}(\widehat{\varepsilon}_{tli}^{(k)})^2) \\
&= r_{tij} + \widehat{\delta}_{tij} + \Delta_{\varepsilon,\beta}^{tij} + \frac{1}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\varepsilon_{tli}^{(k)}\varepsilon_{tlj}^{(k)} + \beta_{tij}(\varepsilon_{tlj}^{(k)})^2 + \beta_{tji}(\varepsilon_{tli}^{(k)})^2) \\
&= \widehat{\delta}_{tij} + \Delta_{\varepsilon,\beta}^{tij} + \frac{1}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\varepsilon_{tli}^{(k)}\varepsilon_{tlj}^{(k)} - r_{tij}) \\
&+ \frac{\beta_{tij}}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}((\varepsilon_{tlj}^{(k)})^2 - r_{tjj}) + \frac{\beta_{tji}}{np}\sum_{k=1}^{n_t}\sum_{l=1}^{p}((\varepsilon_{tli}^{(k)})^2 - r_{tii}) \\
&= \widehat{\delta}_{tij} + \Delta_{\varepsilon,\beta}^{tij} + \widetilde{r}_{tij} - r_{tij} + \beta_{tij}(\widetilde{r}_{tjj} - r_{tjj}) + \beta_{tji}(\widetilde{r}_{tii} - r_{tii}) \\
&= \widehat{\delta}_{tij} + \widetilde{\delta}_{tij} + \Delta_{\varepsilon,\beta}^{tij} + \beta_{tij}\widetilde{\delta}_{tjj} + \beta_{tji}\widetilde{\delta}_{tii}.
\end{aligned} \tag{A.27}$$

The $i = j$ case is trivial, since $\widehat{\rho}_{tii} = \rho_{tii} = 1$. For $i \neq j$, we have

$$\begin{aligned}
\widehat{\rho}_{tij} - \rho_{tij} &= -\frac{\widehat{r}_{tij}\sqrt{r_{tii}r_{tjj}} - r_{tij}\sqrt{\widehat{r}_{tii}\widehat{r}_{tjj}}}{\sqrt{\widehat{r}_{tii}\widehat{r}_{tjj}r_{tii}r_{tjj}}} \\
&= -\frac{(r_{tij} - (\widehat{\delta}_{tij} + \widetilde{\delta}_{tij} + \Delta^{tij}_{\varepsilon,\beta} + \beta^t_{ij}\widetilde{\delta}_{tjj} + \beta^t_{ji}\widetilde{\delta}_{tii}))\sqrt{r_{tii}r_{tjj}}}{\sqrt{\widehat{r}_{tii}\widehat{r}_{tjj}r_{tii}r_{tjj}}} \\
&+ \frac{r_{tij}\sqrt{(r_{tii} + \widetilde{\delta}_{tii} + \widehat{\delta}_{tii})(r_{tjj} + \widetilde{\delta}_{tjj} + \widehat{\delta}_{tjj})}}{\sqrt{\widehat{r}_{tii}\widehat{r}_{tjj}r_{tii}r_{tjj}}} \\
&= \frac{\widehat{\delta}_{tij}\sqrt{r_{tii}r_{tjj}} - \frac{\sqrt{r_{tjj}}}{\sqrt{r_{tii}}}r_{tij}\widetilde{\delta}_{tii} - \frac{\sqrt{r_{tii}}}{\sqrt{r_{tjj}}}r_{tij}\widetilde{\delta}_{tjj} + \frac{\sqrt{r_{tjj}}}{2\sqrt{r_{tii}}}r_{tij}\widetilde{\delta}_{tii} + \frac{\sqrt{r_{tii}}}{2\sqrt{r_{tjj}}}r_{tij}\widetilde{\delta}_{tjj}}{\sqrt{\widehat{r}_{tii}\widehat{r}_{tjj}r_{tii}r_{tjj}}} + O(\Delta_{tij}) \\
&= \frac{\widehat{\delta}_{tij}\sqrt{r_{tii}r_{tjj}} - \frac{\sqrt{r_{tjj}}}{2\sqrt{r_{tii}}}r_{tij}\widetilde{\delta}_{tii} - \frac{\sqrt{r_{tii}}}{2\sqrt{r_{tjj}}}r_{tij}\widetilde{\delta}_{tjj}}{\sqrt{\widehat{r}_{tii}\widehat{r}_{tjj}r_{tii}r_{tjj}}} + O(\Delta_{tij}) \\
&= \frac{\widetilde{\delta}_{tij}}{\sqrt{r_{tii}r_{tjj}}} - \frac{r_{tij}\widetilde{\delta}_{tjj}}{2r_{tjj}\sqrt{r_{tii}r_{tjj}}} - \frac{r_{tij}\widetilde{\delta}_{tii}}{2r_{tii}\sqrt{r_{tii}r_{tjj}}} + O(\Delta'_{tij}),
\end{aligned}$$

where Δ'_{tij} are a collection of high order terms with the same or higher order than $\Delta^{tij}_{\varepsilon,\beta}$, $\widehat{\delta}_{tij}/\widehat{\delta}_{tii}/\widehat{\delta}_{tjj}$, and $\widetilde{\delta}^2_{tij}/\widetilde{\delta}^2_{tii}/\widetilde{\delta}^2_{tjj}/\widetilde{\delta}_{tij}\widetilde{\delta}_{tii}/\widetilde{\delta}_{tij}\widetilde{\delta}_{tjj}$. Therefore

$$\begin{aligned}
&|\sum_{t=1}^{d}\widehat{\rho}_{tij} - \rho_{tij} - \frac{\widetilde{\delta}_{tij}}{\sqrt{r_{tii}r_{tjj}}} + \frac{r_{tij}\widetilde{\delta}_{tjj}}{2r_{tjj}\sqrt{r_{tii}r_{tjj}}} + \frac{r_{tij}\widetilde{\delta}_{tii}}{2r_{tii}\sqrt{r_{tii}r_{tjj}}}| \\
&\qquad = O(\sum_{t=1}^{d}\Delta^{tij}_{\varepsilon,\beta} + \sum_{t=1}^{d}\widehat{\delta}_{tij} + \sum_{t=1}^{d}\widetilde{\delta}^2_{tii}).
\end{aligned}$$

For the first term, notice that $\widehat{\beta}_{tij}$ and $\widehat{\beta}_{tji}$ are bounded as in Theorem 1, with Lemma 7, we have

$$\begin{aligned}
\sum_{t=1}^{d}\Delta^{tij}_{\varepsilon,\beta} &= \sum_{t=1}^{d}\frac{1}{n_t p}\sum_{k=1}^{n_t}\sum_{l=1}^{p}(\widehat{\beta}_{tij}((\widehat{\varepsilon}^{(k)}_{tlj})^2 - (\varepsilon^{(k)}_{tlj})^2) + \widehat{\beta}_{tji}((\widehat{\varepsilon}^{(k)}_{tli})^2 - (\varepsilon^{(k)}_{tli})^2)) \\
&= \sum_{t=1}^{d}(\beta_{tij}\widehat{\delta}_{tjj} + \beta_{tji}\widehat{\delta}_{tii}) + \sum_{t=1}^{d}\Big((\widehat{\beta}_{tij} - \beta_{tij})\widehat{\delta}_{tjj} + (\widehat{\beta}_{tji} - \beta_{tji})\widehat{\delta}_{tii}\Big) \\
&\leq \sum_{t=1}^{d}(\beta_{tij}\widehat{\delta}_{tjj} + \beta_{tji}\widehat{\delta}_{tii}) + \max_{1\leq t\leq d}|\widehat{\beta}_{tij} - \beta_{tij}| \cdot \sum_{t=1}^{d}|\widehat{\delta}_{tjj}| + \max_{1\leq t\leq d}|\widehat{\beta}_{tji} - \beta_{tji}| \cdot \sum_{t=1}^{d}|\widehat{\delta}_{tii}| \\
&= O(\sum_{t=1}^{d}|\widehat{\delta}_{tii}|) + O(\sum_{t=1}^{d}|\widehat{\delta}_{tjj}|) \\
&= O(s\lambda_*^2).
\end{aligned}$$

For the second term, by Lemma 7, we have

$$\sum_{t=1}^{d} \widehat{\delta}_{tij} = O(s\lambda_*^2).$$

For the third term, we have

$$\sum_{t=1}^{d} \widetilde{\delta}_{tjj}^2 \leq \max_{1\leq t\leq d} |\widetilde{\delta}_{tjj}| \cdot \sum_{t=1}^{d} |\widetilde{\delta}_{tjj}| \leq Cd\frac{\log(d/\delta_1)}{n_0 p},$$

where the last inequality holds by further noticing that $\max_{1\leq t\leq d} |\widetilde{\delta}_{tjj}| < \sqrt{C\frac{\log(d/\delta_1)}{n_0 p}}$ with probability $1-\delta_1/k$ by Lemma 1 and Lemma E.1 in [50], where C is a positive constant depending on M only, and $\log(\delta_1^{-1}) = o(n_0)$. Combining all three terms, under Assumption 5, we have

$$\sum_{t=1}^{d} |\widehat{\rho}_{tij} - \rho_{tij} - \frac{\widetilde{\delta}_{tij}}{\sqrt{r_{tii}r_{tjj}}} + \frac{r_{tij}\widetilde{\delta}_{tjj}}{2r_{tjj}\sqrt{r_{tii}r_{tjj}}} + \frac{r_{tij}\widetilde{\delta}_{tii}}{2r_{tii}\sqrt{r_{tii}r_{tjj}}}| = O(s\lambda_*^2 + d\frac{\log(d/\delta_1)}{n_0 p}) = O(s\lambda_*^2),$$

□

Proof of Proposition 2

Proof. Denote by $\boldsymbol{\Theta}_{\mathcal{S}}^P = \frac{1}{\sqrt{d}}\sum_{t=1}^{d}\sqrt{n_t p}\boldsymbol{\xi}_{t\mathcal{S}} \circ \boldsymbol{\Theta}_{t\mathcal{S}}$, and $\boldsymbol{O}_{\mathcal{S}}^P = \frac{1}{\sqrt{d}}\sum_{t=1}^{d}\sqrt{n_t p}\boldsymbol{\xi}_{t\mathcal{S}} \circ \boldsymbol{O}_{t\mathcal{S}}$. For any $x > 0$ and $\delta > 0$, we have

$$\begin{aligned}
&\mathbb{P}(\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty > x) \leq \mathbb{P}(\|\boldsymbol{\Theta}_{\mathcal{S}}^P\|_\infty > x-\delta) + \mathbb{P}(\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty > \delta) \\
&\leq \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x-\delta) + d + \mathbb{P}(\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty > \delta) \\
&= \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x) + \mathbb{P}(x-\delta < \|\boldsymbol{\zeta}\|_\infty \leq x) + d + \mathbb{P}(\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty > \delta),
\end{aligned}$$

where $d = \sup_{x>0} |\mathbb{P}(\|\boldsymbol{\Theta}_S^P\|_\infty > x) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)|$. Similarly,

$$\begin{aligned}
&\mathbb{P}(\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty > x) \geq \mathbb{P}(\|\boldsymbol{\Theta}_{\mathcal{S}}^P\|_\infty > x+\delta) - \mathbb{P}(\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty > \delta) \\
&\geq \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x+\delta) - d - \mathbb{P}(\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty > \delta) \\
&= \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x) - \mathbb{P}(x < \|\boldsymbol{\zeta}\|_\infty \leq x+\delta) - d - \mathbb{P}(\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty > \delta).
\end{aligned}$$

Therefore we conclude that

$$\begin{aligned}
&\sup_{x>0} |\mathbb{P}(\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty > x) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)| \\
&\quad \leq d + \sup_{x>0} \mathbb{P}(x-\delta \leq \|\boldsymbol{\zeta}\|_\infty \leq x+\delta) + \mathbb{P}(\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty > \delta).
\end{aligned} \tag{A.28}$$

By Corollary 1 in [18], it holds that as $\delta \to 0$

$$\sup_{x>0} \mathbb{P}(x-\delta \leq \|\boldsymbol{\zeta}\|_\infty \leq x+\delta) \leq C\delta(\log q)^{1/2}.$$

By Proposition 1, and further notice that $\lambda_* = O(\sqrt{\frac{d+\log q}{n_0 p}})$, we have $\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty = O(\sqrt{\frac{n_0 p}{d}} s\lambda_*^2)$; by Assumption 4 , we have $\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty = o(\frac{1}{\sqrt{\log q}})$. Therefore, we can find $\delta = o(\frac{1}{\sqrt{\log q}})$, such that both $\sup_{x>0} \mathbb{P}(x-\delta < \|\boldsymbol{\zeta}\|_\infty \le x+\delta)$ and $\mathbb{P}(\|\boldsymbol{O}_{\mathcal{S}}^P\|_\infty > \delta)$ goes to 0. For terms in Equation (A.28), it's sufficient to show

$$d = \sup_{x>0} |\mathbb{P}(\|\boldsymbol{\Theta}_{\mathcal{S}}^P\|_\infty > x) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)| \to 0. \tag{A.29}$$

Recall that $\boldsymbol{\Theta}_{\mathcal{S}}^P = \frac{1}{\sqrt{d}} \sum_{t=1}^d \sqrt{n_t p} \boldsymbol{\xi}_{t\mathcal{S}} \circ \boldsymbol{\Theta}_{t\mathcal{S}}$; for each element in $\boldsymbol{\Theta}_{\mathcal{S}}^P$, we rewrite

$$\begin{aligned}
\Theta_{\mathcal{S}}^P(i) &= \frac{1}{\sqrt{d}} \sum_{t=1}^d \sqrt{n_t p} \xi_{t,\chi_1(i),\chi_2(i)} \Theta_{t\mathcal{S}}(i) = \frac{1}{\sqrt{d}} \sum_{t=1}^d \sqrt{n_t p} \xi_{t,\chi_1(i),\chi_2(i)} \Theta_{t\mathcal{S}}(i) \\
&= \frac{1}{\sqrt{d}} \sum_{t=1}^d \frac{\xi_{t,\chi_1(i),\chi_2(i)}}{\sqrt{n_t p}} \sum_{k=1}^{n_t} \sum_{l=1}^{p} \Big(\frac{\varepsilon_{tl\chi_1(i)}^{(k)} \varepsilon_{tl\chi_2(i)}^{(k)} - r_{t,\chi_1(i),\chi_2(i)}^l}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} \\
&\quad - \frac{r_{t,\chi_1(i),\chi_2(i)}}{2 r_{t,\chi_1(i),\chi_1(i)}} \frac{(\varepsilon_{tl\chi_1(i)}^{(k)})^2 - r_{t,\chi_1(i),\chi_1(i)}^l}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} - \frac{r_{t,\chi_1(i),\chi_2(i)}}{2 r_{t,\chi_2(i),\chi_2(i)}} \frac{(\varepsilon_{tl\chi_2(i)}^{(k)})^2 - r_{t,\chi_2(i),\chi_2(i)}^l}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} \Big).
\end{aligned}$$

Denote by $N = \sum_{t=1}^d n_t$, then we have

$$\begin{aligned}
\Theta_{\mathcal{S}}^P(i) &= \frac{1}{\sqrt{Np}} \sum_{t=1}^d \frac{\xi_{t,\chi_1(i),\chi_2(i)} \sqrt{N}}{\sqrt{n_t d}} \sum_{k=1}^{n_t} \sum_{l=1}^{p} \Big(\frac{\varepsilon_{tl\chi_1(i)}^{(k)} \varepsilon_{tl\chi_2(i)}^{(k)} - r_{t,\chi_1(i),\chi_2(i)}^l}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} \\
&\quad - \frac{r_{t,\chi_1(i),\chi_2(i)}}{2 r_{t,\chi_1(i),\chi_1(i)}} \frac{(\varepsilon_{tl\chi_1(i)}^{(k)})^2 - r_{t,\chi_1(i),\chi_1(i)}^l}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} - \frac{r_{t,\chi_1(i),\chi_2(i)}}{2 r_{t,\chi_2(i),\chi_2(i)}} \frac{(\varepsilon_{tl\chi_2(i)}^{(k)})^2 - r_{t,\chi_2(i),\chi_2(i)}^l}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} \Big).
\end{aligned}$$

By Assumption 1, $\frac{\xi_{t,\chi_1(i),\chi_2(i)}\sqrt{N}}{\sqrt{n_t d}}$ is bounded by constant;now applying Lemma 1 to each term in the bracket,

$$\begin{aligned}
\sum_{k=1}^{n_t} \sum_{l=1}^{p} (\varepsilon_{tl\chi_1(i)}^{(k)} \varepsilon_{tl\chi_2(i)}^{(k)} - r_{t,\chi_1(i),\chi_2(i)}^l) &= \sum_{k=1}^{n_t} \sum_{l=1}^{p} \lambda_{tl} (\epsilon_{tl\chi_1(i)}^{(k)} \epsilon_{tl\chi_2(i)}^{(k)} - r_{t,\chi_1(i),\chi_2(i)}), \\
\sum_{k=1}^{n_t} \sum_{l=1}^{p} ((\varepsilon_{tl\chi_1(i)}^{(k)})^2 - r_{t,\chi_1(i),\chi_1(i)}^l) &= \sum_{k=1}^{n_t} \sum_{l=1}^{p} \lambda_{tl} ((\epsilon_{tl\chi_1(i)}^{(k)})^2 - r_{t,\chi_1(i),\chi_1(i)}),
\end{aligned}$$

we can rewrite

$$\Theta_{\mathcal{S}}^P(i) = \frac{1}{\sqrt{Np}} \sum_{t=1}^d \sum_{k=1}^{n_t} \sum_{l=1}^{p} \theta_{\mathcal{S},tl}^{(k)}(i),$$

with

$$\begin{aligned}
\theta_{\mathcal{S},tl}^{(k)}(i) &= \frac{\lambda_{tl} \xi_{t,\chi_1(i),\chi_2(i)} \sqrt{N}}{\sqrt{n_t d}} \Big(\frac{\epsilon_{tl\chi_1(i)}^{(k)} \epsilon_{tl\chi_2(i)}^{(k)} - r_{t,\chi_1(i),\chi_2(i)}}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} \\
&\quad - \frac{r_{t,\chi_1(i),\chi_2(i)}}{2 r_{t,\chi_1(i),\chi_1(i)}} \frac{(\epsilon_{tl\chi_1(i)}^{(k)})^2 - r_{t,\chi_1(i),\chi_1(i)}}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} - \frac{r_{t,\chi_1(i),\chi_2(i)}}{2 r_{t,\chi_2(i),\chi_2(i)}} \frac{(\epsilon_{tl\chi_2(i)}^{(k)})^2 - r_{t,\chi_2(i),\chi_2(i)}^l}{\sqrt{r_{t,\chi_1(i),\chi_1(i)} r_{t,\chi_2(i),\chi_2(i)}}} \Big).
\end{aligned}$$

Now notice that $\epsilon^{(k)}_{tl\chi_1(i)}$ and $\epsilon^{(k)}_{tl\chi_2(i)}$ are Gaussian random variables; therefore we have each element $\theta^{(k)}_{\mathcal{S},tl}(i)$ to be sub-exponential random variable which is independent for over $1 \le t \le d,\ 1 \le l \le p,\ 1 \le k \le n_t$.

According to Corollary 2.1 in [17], our case can be fitted in (E.1): (1)Our random variables are sub-exponential random variables. (2)By Assumption 4, the condition $(\log(n_0dpq))^7/n_0dp = O((n_0pd)^{-c})$ with $c > 0$ is satisfied. We conclude that Equation (A.29) holds with $\sup_{x>0} |\mathbb{P}(\|\Delta \boldsymbol{P}_{\mathcal{S}}\|_\infty > x) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)| = O\left((n_0pd)^{-c}\right)$, and we finish our proof. □

A.4 Proof of Main Theorems

Proof of Theorem 1

Proof. Let $\boldsymbol{D}^{1/2}_{i(l)}\boldsymbol{\beta}^0_{i(l)} = \overline{\boldsymbol{\beta}}^0_{i(l)}$, $\boldsymbol{D}^{1/2}_{i(l)}\widehat{\boldsymbol{\beta}}^0_{i(l)} = \widehat{\overline{\boldsymbol{\beta}}}^0_{i(l)}$, $\boldsymbol{Z}_{t,\cdot,-i}\boldsymbol{D}^{-1/2}_{ti} = \overline{\boldsymbol{Z}}_{t,\cdot,-i}$. We immediately have

$$\begin{aligned}
&\frac{1}{2n_0p}\sum_{t=1}^{d}(\|\boldsymbol{Z}_{t,\cdot,i} - \overline{\boldsymbol{Z}}_{t,\cdot,-i}\widehat{\overline{\boldsymbol{\beta}}}_{ti}\|_2^2 - \|\boldsymbol{Z}_{t,\cdot,i} - \overline{\boldsymbol{Z}}_{t,\cdot,-i}\overline{\boldsymbol{\beta}}_{ti}\|_2^2) \\
&\quad\le \lambda_i\sum_{l\neq i}(\|\overline{\boldsymbol{\beta}}^0_{i(l)}\|_2 - \|\widehat{\overline{\boldsymbol{\beta}}}^0_{i(l)}\|_2) \\
&\quad\le \lambda_i(\sum_{l\in T_i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2 - \sum_{l\in T_i^c}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2),
\end{aligned} \tag{A.30}$$

where $\boldsymbol{\Delta}_i = (\boldsymbol{\Delta}'_{1i}, \boldsymbol{\Delta}'_{2i}, \cdots \boldsymbol{\Delta}'_{di})' = \widehat{\boldsymbol{\beta}}^0_i - \boldsymbol{\beta}^0_i$, $\overline{\boldsymbol{\Delta}}_i = \boldsymbol{D}^{1/2}_i\boldsymbol{\Delta}_i = \widehat{\overline{\boldsymbol{\beta}}}^0_i - \overline{\boldsymbol{\beta}}^0_i$, and $\overline{\boldsymbol{\Delta}}_{i(l)}$ is the l-th group of $\overline{\boldsymbol{\Delta}}_i$; $T_i = \{l : \boldsymbol{\beta}^0_{i(l)} \neq 0\} = \{l : \overline{\boldsymbol{\beta}}^0_{i(l)} \neq 0\}$ is the group support of vector $\boldsymbol{\beta}^0_i$, while T_i^c is its complement set. By the convexity of ℓ_2 norm, we have on event E_i,

$$\begin{aligned}
&\frac{1}{2n_0p}\sum_{t=1}^{d}(\|\boldsymbol{Z}_{t,\cdot,i} - \overline{\boldsymbol{Z}}_{t,\cdot,-i}\widehat{\overline{\boldsymbol{\beta}}}_{ti}\|_2^2 - \|\boldsymbol{Z}_{t,\cdot,i} - \overline{\boldsymbol{Z}}_{t,\cdot,-i}\overline{\boldsymbol{\beta}}_{ti}\|_2^2) \\
&\quad\ge \frac{1}{n_0p}\sum_{t=1}^{d}\overline{\Delta}'_{ti}\overline{\boldsymbol{Z}}'_{t,\cdot,-i}\boldsymbol{\varepsilon}_{ti} \\
&\quad\ge -\sum_{l\neq i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2 \cdot \max_{l\neq i}\frac{[\sum_{t=1}^{d}(\overline{\boldsymbol{Z}}'_{t,\cdot,l}\boldsymbol{\varepsilon}_{ti})^2]^{1/2}}{n_0p} \\
&\quad\ge -\frac{\xi-1}{\xi+1}\lambda_i(\sum_{l\neq i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2)
\end{aligned} \tag{A.31}$$

where the last inequality follows from Lemma 3 on event E_i. Combining Equation (A.30) and Equation (A.31), we have

$$\lambda_i(\sum_{l\in T_i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2 - \sum_{l\in T_i^c}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2) \ge -\frac{\xi-1}{\xi+1}\lambda_i(\sum_{l\neq i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2),$$

which further implies that on E_i,

$$\sum_{l\in T_i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2 \leq \xi\sum_{l\in T_i^c}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2. \tag{A.32}$$

Now, also notice that

$$\begin{aligned}\frac{1}{2n_0p}\sum_{t=1}^{d}(\|\boldsymbol{Z}_{t,\cdot,i}-\overline{\boldsymbol{Z}}_{t,\cdot,-i}\widehat{\overline{\boldsymbol{\beta}}}_{ti}\|_2^2-\|\boldsymbol{Z}_{t,\cdot,i}-\overline{\boldsymbol{Z}}_{t,\cdot,-i}\overline{\boldsymbol{\beta}}_{ti}\|_2^2)\\ =\frac{1}{2n_0p}\sum_{t=1}^{d}(\|\overline{\boldsymbol{Z}}_{t,\cdot,-i}\overline{\boldsymbol{\Delta}}_{ti}\|_2^2-2\overline{\boldsymbol{\Delta}}_{ti}^{'}\overline{\boldsymbol{Z}}_{t,\cdot,-i}^{'}\boldsymbol{E}_{ti}).\end{aligned} \tag{A.33}$$

In addition, by Lemma 1, we have on event E_i,

$$\frac{1}{n_0p}|\sum_{t=1}^{d}\overline{\boldsymbol{\Delta}}_{ti}^{'}\overline{\boldsymbol{Z}}_{t,\cdot,-i}^{'}\boldsymbol{E}_{ti})| \leq \sum_{l\neq i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2\cdot\max_{l\neq i}\frac{[\sum_{t=1}^{d}(\overline{\boldsymbol{Z}}_{t,\cdot,l}^{'}\boldsymbol{E}_{ti})^2]^{1/2}}{n_0p}\leq\lambda_i\frac{\xi-1}{\xi+1}\sum_{l\neq i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2, \tag{A.34}$$

where the last inequality follows from Lemma 3 on E_i. Combining Equation (A.30), Equation (A.33) and Equation (A.34), we obtain that on E_i,

$$\begin{aligned}\frac{1}{2n_0p}\sum_{t=1}^{d}\|\overline{\boldsymbol{Z}}_{t,\cdot,-i}\overline{\boldsymbol{\Delta}}_{ti}\|_2^2 &\leq \lambda_i(\sum_{l\in T_i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2-\sum_{l\in T_i^c}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2)+\lambda_i\frac{\xi-1}{\xi+1}(\sum_{l\in T_i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2+\sum_{l\in T_i^c}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2)\\ &\leq \lambda_i\frac{2\xi}{\xi+1}\sum_{l\in T_i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2.\end{aligned} \tag{A.35}$$

On the other hand, we have

$$\begin{aligned}\frac{1}{2n_0p}\sum_{t=1}^{d}\|\overline{\boldsymbol{Z}}_{t,\cdot,-i}\overline{\boldsymbol{\Delta}}_{ti}\|_2^2 &= \frac{1}{2n_0p}\sum_{t=1}^{d}\|\boldsymbol{Z}_{t,\cdot,-i}\boldsymbol{D}_{ti}^{-1/2}\overline{\boldsymbol{\Delta}}_{ti}\|_2^2\\ &=\frac{1}{2}\sum_{t=1}^{d}[\overline{\boldsymbol{\Delta}}_{ti}^{'}\boldsymbol{D}_{ti}^{-1/2}(\frac{\boldsymbol{Z}_{t,\cdot,-i}^{'}\boldsymbol{Z}_{t,\cdot,-i}}{n_0p}-\overline{\boldsymbol{\Psi}}_{ti})\boldsymbol{D}_{ti}^{-1/2}\overline{\boldsymbol{\Delta}}_{ti}\\ &+\overline{\boldsymbol{\Delta}}_{ti}^{'}\boldsymbol{D}_{ti}^{-1/2}\overline{\boldsymbol{\Psi}}_{ti}\boldsymbol{D}_{ti}^{-1/2}\overline{\boldsymbol{\Delta}}_{ti}],\end{aligned} \tag{A.36}$$

where $\overline{\boldsymbol{\Psi}}_{ti}=\mathbb{E}\frac{\boldsymbol{Z}_{t,\cdot,-i}^{'}\boldsymbol{Z}_{t,\cdot,-i}}{n_0p}=\frac{\mathbb{E}\boldsymbol{X}_{t,\cdot,-i}^{(1)}{}^{'}\boldsymbol{X}_{t,\cdot,-i}^{(1)}}{p}\cdot\frac{n_t}{n_0}=\boldsymbol{V}_{t,-i,-i}\frac{n_t}{n_0}$ by Assumption 2. By Lemma 4, we know that on E_{dia}, all diagonal terms of $\boldsymbol{D}_{ti}$ for all $1\leq t\leq d$ are bounded above and below, where the event E_{dia} hold with probability $1-(dq)^{-\delta}$ as

$$E_{dia}=\{|\frac{\boldsymbol{Z}_{t,\cdot,i}^{'}\boldsymbol{Z}_{t,\cdot,i}}{n_0p}|\in(\frac{1}{2c_e},2c_e)\text{ for all }1\leq t\leq d\text{ and }1\leq i\in q\}.$$

Hence, by Assumption 1 and Assumption 3, we can bound the second term on the right hand side of Equation (A.36) as

$$\sum_{t=1}^{d} \overline{\boldsymbol{\Delta}}_{ti}^{'} \boldsymbol{D}_{ti}^{-1/2} \overline{\boldsymbol{\Psi}}_{ti} \boldsymbol{D}_{ti}^{-1/2} \overline{\boldsymbol{\Delta}}_{ti} \geq \min_{1\leq t\leq d} \lambda_{min}(\overline{\boldsymbol{\Psi}}_{ti}) \|\overline{\boldsymbol{\Delta}}_i\|_2^2 \frac{1}{2c_e} \geq \frac{1}{2c_e^2} \|\overline{\boldsymbol{\Delta}}_i\|_2^2. \tag{A.37}$$

To handle the first term on the right hand side of Equation (A.36), we notice that

$$|\sum_{t=1}^{d} \overline{\boldsymbol{\Delta}}_{ti}^{'} \boldsymbol{D}_{ti}^{-1/2} (\frac{\boldsymbol{Z}_{t,\cdot,-i}^{'} \boldsymbol{Z}_{t,\cdot,-i}}{n_0 p} - \overline{\boldsymbol{\Psi}}_{ti}) \boldsymbol{D}_{ti}^{-1/2} \overline{\boldsymbol{\Delta}}_{ti}| \leq 2c_e \max_{l_1\neq i, l_2\neq i} \|\boldsymbol{M}_{l_1,l_2}\|_2 \cdot (\sum_{l\neq i} \|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2)^2, \tag{A.38}$$

where the last inequality follows from the proof in norm compression inequality (Theorem 3.4 of [14]) and a simple fact that for any vector $\boldsymbol{v} \in \mathbb{R}^p$ and $\boldsymbol{M} \in \mathbb{R}^{p\times p}$, $\boldsymbol{v}^{'}\boldsymbol{M}\boldsymbol{v} \leq \|\boldsymbol{v}\|_1^2 |\boldsymbol{M}|_\infty$. Here $\boldsymbol{M}_{l_1,l_2} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{M}_{l_1,l_2} = (m_{l_1,l_2,t_1,t_2})$, where $m_{l_1,l_2,t_1,t_2} = \frac{\boldsymbol{Z}_{t_1,\cdot,l_1}^{'} \boldsymbol{Z}_{t_2,\cdot,l_2}}{n_0 p} - \mathbb{E}\frac{\boldsymbol{Z}_{t_1,\cdot,l_1}^{'} \boldsymbol{Z}_{t_2,\cdot,l_2}}{n_0 p}$.

By Lemma 5, we have on event E_{com}, $\max_{l_1\neq i, l_2\neq i} \|M_{l_1,l_2}\|_2 \leq c\sqrt{\frac{d+\log q}{n_0 p}}$, which, together with Equation (A.38), implies that on $E_{com} \cap E_{dia}$

$$\begin{aligned}
&|\sum_{t=1}^{d} \overline{\boldsymbol{\Delta}}_{ti}^{'} \boldsymbol{D}_{ti}^{-1/2} (\frac{\boldsymbol{Z}_{t,\cdot,-i}^{'} \boldsymbol{Z}_{t,\cdot,-i}}{n_0 p} - \overline{\boldsymbol{\Psi}}_{ti}) \boldsymbol{D}_{ti}^{-1/2} \overline{\boldsymbol{\Delta}}_{ti}| \\
&\leq 2c_e c \sqrt{\frac{d+\log q}{n_0 p}} (\sum_{l\in T_i} \|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2 + \sum_{l\in T_i^c} \|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2)^2 \\
&\leq 2c_e c (1+\xi)^2 \sqrt{\frac{d+\log q}{n_0 p}} (\sum_{l\in T_i} \|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2)^2 \\
&\leq 2c_e c (1+\xi)^2 \sqrt{\frac{d+\log q}{n_0 p}} s \|\overline{\boldsymbol{\Delta}}_i\|_2^2,
\end{aligned} \tag{A.39}$$

where the last two steps follow from Equation (A.32) and Cauchy–Schwarz inequality, respectively. Combining Equation (A.36)-(A.39), we have on $E_{com\cap E_{dia}}$,

$$\frac{1}{2n_0 p} \sum_{t=1}^{d} \|\overline{\boldsymbol{Z}}_{t,\cdot,-i} \overline{\boldsymbol{\Delta}}_{ti}\|_2^2 \geq \frac{1}{8c_e^2} \|\overline{\boldsymbol{\Delta}}_i\|_2, \tag{A.40}$$

where we have used Assumption 4 which implies $s\sqrt{\frac{d+\log q}{n_0 p}} = o(1)$. Define $E_{joint,i} = E_{com} \cap E_{dia} \cap E_i$, finally combing Equation (A.35) and Equation (A.40), on $E_{joint,i}$,

$$\|\overline{\boldsymbol{\Delta}}_i\|_2^2 \leq 8c_e^2 \lambda_i \frac{2\xi}{\xi+1} \sum_{l\in T_i} \|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2 \leq 8c_e^2 \lambda_i \frac{2\xi}{\xi+1} \sqrt{s} \|\overline{\boldsymbol{\Delta}}_i\|_2,$$

which implies

$$\|\overline{\boldsymbol{\Delta}}_i\|_2 \leq C\sqrt{s\frac{d+\log q}{n_0 p}}. \tag{A.41}$$

In addition, on the same event, $E_{joint,i}$

$$\sum_{l\neq i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2 \leq (1+\xi)\sum_{l\in T_i}\|\overline{\boldsymbol{\Delta}}_{i(l)}\|_2 \leq (1+\xi)\sqrt{s}\|\overline{\boldsymbol{\Delta}}_i\|_2$$
$$\leq C's\sqrt{\frac{d+\log q}{n_0 p}}. \tag{A.42}$$

Moreover, by Equation (A.35) and (A.42),

$$\frac{1}{2n_0p}\sum_{t=1}^{d}\|\overline{\boldsymbol{Z}}_{t,\cdot,-i}\overline{\boldsymbol{\Delta}}_{ti}\|_2^2 \leq \lambda_i\frac{4\xi}{\xi+1}C's\sqrt{\frac{d+\log q}{n_0p}} \leq C''s\frac{d+\log q}{n_0p}. \tag{A.43}$$

Finally, note that on E_{dia}, all entries of $\boldsymbol{D}_{ti}$ are simultaneously bounded below and above by $(\frac{1}{2c_e}, 2c_e)$. In addition, $E_{joint,i}$ holds with probabilty at least $1-C_i$ where $C_i = C_1+C_2+C_3$, i.e, the sum of the constants corresponding to E_{com}, E_{dia} and E_i. Therefore Equation (A.41) - (A.43) directly imply our statement of Theorem 1. □

Proof of Theorem 4

Proof. Our procedure for temporal covariance estimation is applied to each sub-graph separately, therefore we omit index t here for clarity. We split our proof into algebraic and probabilistic parts.

We start by discussing the probability of certain events. Firstly, we notice that $\boldsymbol{I}-\boldsymbol{C}$ and $\boldsymbol{I}-\widehat{\boldsymbol{C}}$ are lower triangular matrix with all diagonal elements as 1, thus it holds naturally that for some $c>0$,

$$\|\boldsymbol{I}-\boldsymbol{C}\|_2 \in (\frac{1}{c},c),\ \|\boldsymbol{I}-\widehat{\boldsymbol{C}}\|_2 \in (\frac{1}{c},c),\ \|(\boldsymbol{I}-\boldsymbol{C})^{-1}\|_2 \in (\frac{1}{c},c),\ \|(\boldsymbol{I}-\widehat{\boldsymbol{C}})^{-1}\|_2 \in (\frac{1}{c},c). \tag{A.44}$$

Next, by Lemma B.2 as in [40] and Assumption 3, we have that for some $c'>0$,

$$\|\frac{q}{\mathrm{Tr}(\boldsymbol{V})}\boldsymbol{D}^{-1}\|_2 \in (\frac{2}{c},\frac{c}{2}),\ \|\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2 \in (\frac{2}{c},\frac{c}{2}).$$

Notice that

$$\|\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2 - \|\widehat{\boldsymbol{D}}-\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2 \leq \|\widehat{\boldsymbol{D}}\|_2 \leq \|\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2 + \|\widehat{\boldsymbol{D}}-\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2,$$

by noticing that $\boldsymbol{D}-\widehat{\boldsymbol{D}}$ is a diagonal matrix, it directly implies that,

$$\|\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2 - \max_{1\leq i\leq p}|\widehat{d}_i-\frac{\mathrm{Tr}(\boldsymbol{V})}{q}d_i| \leq \|\widehat{\boldsymbol{D}}\|_2 \leq \|\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2 + \max_{1\leq i\leq p}|\widehat{d}_i-\frac{\mathrm{Tr}(\boldsymbol{V})}{q}d_i|.$$

By Lemma 10 we can bound $\max_{1\leq i\leq p}|\widehat{d}_i-\frac{\mathrm{Tr}(\boldsymbol{V})}{q}d_i| \leq \frac{1}{c}$ with probability at least $1-q^{-\delta}$; thus we can directly show that with at least the same probability, the event E_d holds for some constant c, where

$$E_d = \{\|\widehat{\boldsymbol{D}}\|_2 \in (\frac{1}{c},c),\ \|\widehat{\boldsymbol{D}}^{-1}\|_2 \in (\frac{1}{c},c),\ \|\frac{q}{\mathrm{Tr}(\boldsymbol{V})}\boldsymbol{D}^{-1}\|_2 \in (\frac{1}{c},c),\ \|\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2 \in (\frac{1}{c},c)\}$$

Now we proceed to algebraic part. Under event E_d, for $\widehat{\boldsymbol{D}}$, we have

$$\frac{1}{p}\|\widehat{\boldsymbol{D}} - \frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_F^2 = \frac{1}{p}\sum_{i=1}^{p}|\widehat{d_i} - \frac{\text{Tr}(\boldsymbol{V})}{q}d_i|^2, \tag{A.45}$$

$$\begin{aligned}\frac{1}{p}\|\widehat{\boldsymbol{D}}^{-1} - (\frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D})^{-1}\|_F^2 &\leq \frac{1}{p}\|\widehat{\boldsymbol{D}} - \frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_F^2 \cdot \|(\frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D})^{-1}\|_2^2 \cdot \|(\widehat{\boldsymbol{D}})^{-1}\|_2^2 \\ &\leq C\frac{1}{p}\|\widehat{\boldsymbol{D}} - \frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_F^2 = C\frac{1}{p}\sum_{i=1}^{p}|\widehat{d_i^*} - \frac{\text{Tr}(\boldsymbol{V})}{q}d_i|^2, \end{aligned}\tag{A.46}$$

For $\widehat{\boldsymbol{C}}$,

$$\frac{1}{p}\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_F^2 = \frac{1}{p}\sum_{i=1}^{p}\|\widehat{\boldsymbol{c}}_i^* - \boldsymbol{c}_i\|_2^2. \tag{A.47}$$

$$\begin{aligned}\frac{1}{p}\|(\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1} - (\boldsymbol{I} - \boldsymbol{C})^{-1}\|_F^2 &\leq \frac{1}{p}\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_F^2 \cdot \|(\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1}\|_2^2 \cdot \|(\boldsymbol{I} - \boldsymbol{C})^{-1}\|_2^2 \\ &\leq C\frac{1}{p}\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_F^2 = C\frac{1}{p}\sum_{i=1}^{p}\|\widehat{\boldsymbol{c}}_i^* - \boldsymbol{c}_i\|_2^2, \end{aligned}\tag{A.48}$$

where the last inequality follows from Equation (A.44). Finally, as we defined, $\widehat{\boldsymbol{U}} = (\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1}\widehat{\boldsymbol{D}}((\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1})'$, $\widehat{\boldsymbol{A}} = (\boldsymbol{I} - \widehat{\boldsymbol{C}})'\widehat{\boldsymbol{D}}^{-1}(\boldsymbol{I} - \widehat{\boldsymbol{C}})$, therefore we have

$$\begin{aligned}\frac{1}{p}\|\widehat{\boldsymbol{U}} - \frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{U}\|_F^2 \leq \frac{3}{p}\Bigg(&\|(\boldsymbol{I} - \boldsymbol{C})^{-1}\|_2^2 \cdot \|\frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_2^2 \cdot \|(\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1} - (\boldsymbol{I} - \boldsymbol{C})^{-1}\|_F^2 \\ &+ \|(\boldsymbol{I} - \boldsymbol{C})^{-1}\|_2^2 \cdot \|\widehat{\boldsymbol{D}} - \frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D}\|_F^2 \cdot \|(\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1}\|_2^2 \\ &+ \|(\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1} - (\boldsymbol{I} - \boldsymbol{C})^{-1}\|_F^2 \cdot \|\widehat{\boldsymbol{D}}\|_2^2 \cdot \|(\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1}\|_2^2\Bigg).\end{aligned}$$

$$\begin{aligned}\frac{1}{p}\|\widehat{\boldsymbol{A}} - \frac{q}{\text{Tr}(\boldsymbol{V})}\boldsymbol{A}\|_F^2 \leq \frac{3}{p}\Bigg(&\|\boldsymbol{I} - \boldsymbol{C}\|_2^2 \cdot \|\frac{q}{\text{Tr}(\boldsymbol{V})}\boldsymbol{D}^{-1}\|_2^2 \cdot \|\boldsymbol{C} - \widehat{\boldsymbol{C}}\|_F^2 \\ &+ \|\boldsymbol{I} - \boldsymbol{C}\|_2^2 \cdot \|\widehat{\boldsymbol{D}}^{-1} - (\frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D})^{-1}\|_F^2 \cdot \|\boldsymbol{I} - \widehat{\boldsymbol{C}}\|_2^2 \\ &+ \|\boldsymbol{C} - \widehat{\boldsymbol{C}}\|_F^2 \cdot \|\widehat{\boldsymbol{D}}^{-1}\|_2^2 \cdot \|\boldsymbol{I} - \widehat{\boldsymbol{C}}\|_2^2\Bigg).\end{aligned}$$

Notice that under E_i and by Equation (A.44), $\|\frac{q}{\text{Tr}(\boldsymbol{V})}\widehat{\boldsymbol{D}}^{-1}\|_2$, $\|\frac{\text{Tr}(\boldsymbol{V})}{q}\widehat{\boldsymbol{D}}\|_2$, $\|(\boldsymbol{I}-\boldsymbol{C})^{-1}\|_2$, $\|\boldsymbol{I}-\boldsymbol{C}\|_2$, $\|(\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1}\|_2$ and $\|\boldsymbol{I} - \widehat{\boldsymbol{C}}\|_2$ are bounded by a constant, thus we have

$$\frac{1}{p}\|\widehat{\boldsymbol{U}} - \frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{U}\|_F^2 \leq \frac{C_1}{p}\|\widehat{\boldsymbol{D}} - \frac{\text{Tr}(\boldsymbol{V})}{q}\boldsymbol{D})\|_F^2 + \frac{C_2}{p}\|(\boldsymbol{I} - \widehat{\boldsymbol{C}})^{-1} - (\boldsymbol{I} - \boldsymbol{C})^{-1}\|_F^2.$$

$$\frac{1}{p}\|\widehat{\boldsymbol{A}}-\frac{q}{\mathrm{Tr}(\boldsymbol{V})}\boldsymbol{A}\|_F^2\leq\frac{C_1}{p}\|\widehat{\boldsymbol{D}}^{-1}-(\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{D})^{-1}\|_F^2+\frac{C_2}{p}\|\widehat{\boldsymbol{C}}-\boldsymbol{C}\|_F^2.$$

With Lemma 10, Equation (A.45)- Equation (A.48), we show that for for any $\delta>0$, there exists constants C, such that

$$\mathbb{P}(\frac{1}{p}\|\widehat{\boldsymbol{U}}-\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{U}\|_F^2\geq C\frac{\log p}{(nq)^{\frac{2\alpha+1}{2\alpha+2}}})\leq q^{-\delta},$$

$$\mathbb{P}(\frac{1}{p}\|\widehat{\boldsymbol{A}}-\frac{q}{\mathrm{Tr}(\boldsymbol{V})}\boldsymbol{A}\|_F^2\geq C\frac{\log p}{(nq)^{\frac{2\alpha+1}{2\alpha+2}}})\leq q^{-\delta}.$$

And further notice that E_d holds with probability at least $1-q^{-\delta}$, thus we proved the main conclusion. □

Proof of Corollary 2

Proof. Following the proof of Theorem 4 and omitting index t, we denote by $\widetilde{\boldsymbol{U}}=\frac{\mathrm{Tr}(\boldsymbol{V})}{q}\boldsymbol{U}$; we denote by $J(\widehat{\boldsymbol{U}})$, the set of non-zero indices in $\widehat{\boldsymbol{U}}$ by $\widehat{J}$; we denote by $\widehat{J}^c$ as its complement set. We immediately have $|\widehat{J}|\leq p(2h+1)$. By Cauchy-Schwarz inequality,

$$\left(\sum_{(i,j)\in\widehat{J}}|\widehat{U}_{ij}-\widetilde{U}_{ij}|\right)^2\leq|\widehat{J}|\sum_{(i,j)\in\widehat{J}}(\widehat{U}_{ij}-\widetilde{U}_{ij})^2\leq p(2h+1)\|\widehat{\boldsymbol{U}}_{\widehat{J}}-\widetilde{\boldsymbol{U}}_{\widehat{J}}\|_F^2.$$

By Theorem 4, we known that $\|\widehat{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}\|_F^2=O(p\frac{\log p}{(nq)^{\frac{2\alpha+1}{2\alpha+2}}})$, thus

$$\sum_{(i,j)\in\widehat{J}}|\widehat{U}_{ij}-\widetilde{U}_{ij}|=O(\sqrt{p^2h\frac{\log p}{(nq)^{\frac{2\alpha+1}{2\alpha+2}}}})=O(\sqrt{p^2\frac{\log p}{(nq)^{\frac{\alpha}{\alpha+1}}}}).\tag{A.49}$$

Therefore,

$$\begin{aligned}
\left|\|\widehat{\boldsymbol{U}}\|_F^2-\|\widetilde{\boldsymbol{U}}\|_F^2\right|&=\left|\sum_{i=1}^p\sum_{j=1}^p(\widehat{U}_{ij}^2-\widetilde{U}_{ij}^2)\right|=\left|\sum_{(i,j)\in\widehat{J}}(\widehat{U}_{ij}^2-\widetilde{U}_{ij}^2)-\sum_{(i,j)\in\widehat{J}^c}\widetilde{U}_{ij}^2\right|\\
&\leq\left|\sum_{(i,j)\in\widehat{J}}(\widehat{U}_{ij}^2-\widetilde{U}_{ij}^2)\right|+\sum_{(i,j)\in\widehat{J}^c}\widetilde{U}_{ij}^2\\
&=\left|\sum_{(i,j)\in\widehat{J}}(\widehat{U}_{ij}-\widetilde{U}_{ij})(\widehat{U}_{ij}+\widetilde{U}_{ij})\right|+\sum_{(i,j)\in\widehat{J}^c}\widetilde{U}_{ij}^2\\
&\leq\sum_{(i,j)\in\widehat{J}}|\widehat{U}_{ij}-\widetilde{U}_{ij}|(|\widehat{U}_{ij}-\widetilde{U}_{ij}|+2|\widetilde{U}_{ij}|)+\sum_{(i,j)\in\widehat{J}^c}\widetilde{U}_{ij}^2\\
&\leq\|\widehat{\boldsymbol{U}}-\widetilde{\boldsymbol{U}}\|_F^2+2\max_{i,j}(|\widetilde{U}_{ij}|)\sum_{(i,j)\in\widehat{J}}|\widehat{U}_{ij}-\widetilde{U}_{ij}|.
\end{aligned}$$

By Assumption 3, Theorem 4 and Equation (A.49),

$$\left|\|\widehat{\boldsymbol{U}}\|_F^2-\|\widetilde{\boldsymbol{U}}\|_F^2\right| = O\left(p\frac{\log p}{(nq)^{\frac{2\alpha+1}{2\alpha+2}}}+p\sqrt{\frac{\log p}{(nq)^{\frac{\alpha}{\alpha+1}}}}\right) = O\left(p\sqrt{\frac{\log p}{(nq)^{\frac{\alpha}{\alpha+1}}}}\right).$$

By Assumption 6, we have

$$\frac{\|\widehat{\boldsymbol{U}}\|_F^2-\|\widetilde{\boldsymbol{U}}\|_F^2}{p} = o(1), \tag{A.50}$$

With Assumption 2, denote by $T_v = \frac{\mathrm{Tr}(\boldsymbol{V})}{q} = \frac{\mathrm{Tr}(\widetilde{\boldsymbol{U}})}{p}$; by definition of $\widehat{T}_v$ and Equation (A.49), we have

$$\widehat{T}_v - T_v = \frac{\mathrm{Tr}(\widehat{\boldsymbol{U}}) - \mathrm{Tr}(\widetilde{\boldsymbol{U}})}{p} = O\left(\sqrt{\frac{\log p}{(nq)^{\frac{\alpha}{\alpha+1}}}}\right) = o(1). \tag{A.51}$$

Finally,

$$\begin{aligned}
\frac{\|\widehat{\boldsymbol{U}}^*\|_F^2-\|\boldsymbol{U}\|_F^2}{p} &= \frac{\|\widehat{\boldsymbol{U}}\|_F^2-\widehat{T}_v^2\|\boldsymbol{U}\|_F^2}{p\widehat{T}_v^2} = \frac{\|\widehat{\boldsymbol{U}}\|_F^2-\|\widetilde{\boldsymbol{U}}\|_F^2+\|\widetilde{\boldsymbol{U}}\|_F^2-\widehat{T}_v^2\|\boldsymbol{U}\|_F^2}{p\widehat{T}_v^2} \\
&= \frac{\|\widehat{\boldsymbol{U}}\|_F^2-\|\widetilde{\boldsymbol{U}}\|_F^2}{p\widehat{T}_v^2} + \frac{(T_v^2-\widehat{T}_v^2)\|\boldsymbol{U}\|_F^2}{p\widehat{T}_v^2} \\
&\le \frac{1}{\widehat{T}_v^2}\left(\frac{\|\widehat{\boldsymbol{U}}\|_F^2-\|\widetilde{\boldsymbol{U}}\|_F^2}{p} + c_e^2(T_v^2-\widehat{T}_v^2)\right)
\end{aligned}$$

Therefore, combining Equation (A.50) and (A.51), we have $\frac{\|\widehat{\boldsymbol{U}}^*\|_F^2-\|\boldsymbol{U}\|_F^2}{p} = o(1)$. and we finish our proof.

□

Proof of Theorem 2

Proof. Denote $d_\zeta = \sup_{x>0}|\mathbb{P}(\|\widehat{\boldsymbol{\zeta}}\|_\infty > x|\mathcal{D}) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)|$,

$$\begin{aligned}
&\sup_{x>0}|\mathbb{P}(\|\Delta\boldsymbol{P}_S\|_\infty > x) - \mathbb{P}(\|\widehat{\boldsymbol{\zeta}}\|_\infty > x|\mathcal{D})| \\
&\le \sup_{x>0}|\mathbb{P}(\|\Delta\boldsymbol{P}_S\|_\infty > x) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)| + \sup_{x>0}|\mathbb{P}(\|\widehat{\boldsymbol{\zeta}}\|_\infty > x|\mathcal{D}) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)| \\
&= \sup_{x>0}|\mathbb{P}(\|\Delta\boldsymbol{P}_S\|_\infty > x) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)| + d_\zeta.
\end{aligned}$$

From Proposition 2, the first term converges to 0, therefore we only need to prove $d_\zeta \to 0$. Now with Lemma 3.1 in [17],

$$\sup_{x>0}|\mathbb{P}(\|\widehat{\boldsymbol{\zeta}}\|_\infty > x|\mathcal{D}) - \mathbb{P}(\|\boldsymbol{\zeta}\|_\infty > x)| \le C\|\widehat{\boldsymbol{W}}^P - \boldsymbol{W}^P\|_\infty^{1/3}\{1 \vee \log(q/\|\widehat{\boldsymbol{W}}^P - \boldsymbol{W}^P\|_\infty)\}^{2/3}.$$

Therefore it's sufficient to show

$$\|\widehat{\boldsymbol{W}}^P - \boldsymbol{W}^P\|_\infty = o_p(1). \tag{A.52}$$

Proving such a claim require us to calculate each entry in $\boldsymbol{W}^P$. As defined in Equation (2.16),

$$\boldsymbol{W}^P = \mathbb{E}\{\frac{1}{d}\sum_{t=1}^{d}(\sqrt{n_t p}\boldsymbol{\Theta}_{tS}^{\rho})(\sqrt{n_t p}\boldsymbol{\Theta}_{tS}^{\rho})'\} = \frac{1}{d}\sum_{t=1}^{d}\boldsymbol{W}_t^{\rho}.$$

Therefore, $\boldsymbol{W}^P$ follows from $\boldsymbol{W}_t^{\rho}$ immediately. By Lemma 11, we have

$$\begin{aligned}
w_{tij}^P = \frac{1}{d}\sum_{t=1}^{d}\big(&\frac{\|\boldsymbol{U}_t\|_F^2}{p}(\rho_{t,\chi_1(i),\chi_1(j)}\rho_{t,\chi_2(i),\chi_2(j)} + \rho_{t,\chi_1(i),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(i)} \\
&+ \frac{1}{2}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(j),\chi_2(j)}\rho_{t,\chi_1(i),\chi_1(j)}^2 + \frac{1}{2}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(j),\chi_2(j)}\rho_{t,\chi_2(i),\chi_2(j)}^2 \\
&+ \frac{1}{2}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(j),\chi_2(j)}\rho_{t,\chi_1(i),\chi_2(j)}^2 + \frac{1}{2}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(j),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(i)}^2 \\
&- \rho_{t,\chi_1(i),\chi_1(j)}\rho_{t,\chi_1(j),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(i)} - \rho_{t,\chi_1(i),\chi_1(j)}\rho_{t,\chi_1(i),\chi_2(i)}\rho_{t,\chi_1(i),\chi_2(j)} \\
&- \rho_{t,\chi_2(i),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(j)}\rho_{t,\chi_1(i),\chi_2(j)} - \rho_{t,\chi_2(i),\chi_2(j)}\rho_{t,\chi_1(j),\chi_2(i)}\rho_{t,\chi_1(i),\chi_2(i)})\big)
\end{aligned}$$

For $\frac{\|\widehat{\boldsymbol{U}_t}\|_F^2 - \|\boldsymbol{U}_t\|_F^2}{p}$, we know from Corollary 2 that

$$\frac{\|\widehat{\boldsymbol{U}_t}\|_F^2 - \|\boldsymbol{U}_t\|_F^2}{p} = o(1). \tag{A.53}$$

For the plug-in estimator defined in Equation (2.9), we notice that uniformly for arbitrary $i, i', i'', i''', j, j', j'', j'''$, we have

$$\frac{\sum_{t=1}^{d}(\widehat{\rho}_{tij}\widehat{\rho}_{ti'j'} - \rho_{tij}\rho_{ti'j'})}{d} = o(1), \tag{A.54}$$

$$\frac{\sum_{t=1}^{d}(\widehat{\rho}_{tij}\widehat{\rho}_{ti'j'}\widehat{\rho}_{ti''j''} - \rho_{tij}\rho_{ti'j'}\rho_{ti''j''})}{d} = o(1), \tag{A.55}$$

$$\frac{\sum_{t=1}^{d}(\widehat{\rho}_{tij}\widehat{\rho}_{ti'j'}\widehat{\rho}_{ti''j''}\widehat{\rho}_{ti'''j'''} - \rho_{tij}\rho_{ti'j'}\rho_{ti''j''}\rho_{ti'''j'''})}{d} = o(1), \tag{A.56}$$

from Proposition 1 and Lemma 8.

From Equation (A.53), Equation (A.54), Equation (A.55) and Equation (A.56) , we know Equation (A.52) holds, and we finish our proof. □

Proof of Theorem 3

Proof. The first claim directly follows from Theorem 2, and we know that $\mathbb{P}_{H_0}(\boldsymbol{c} \notin C_S(1-\alpha)) \to \alpha$.

For the second claim, we define that $\mu_{\widehat{\boldsymbol{\zeta}}} = \mathbb{E}(\|\widehat{\boldsymbol{\zeta}}\|_\infty|\mathcal{D})$, according to standard theory for the maximum of Gaussian variables, we have

$$\mathbb{P}(\|\widehat{\boldsymbol{\zeta}}\|_\infty \geq \mu_{\widehat{\boldsymbol{\zeta}}} + u|\mathcal{D}) \leq \exp(-\frac{u^2}{2\max_{1\leq j\leq r}\widehat{w}_{jj}^P}).$$

Furthermore, define $C_0 = \sqrt{2}(1 + \frac{1}{2\log q})$,

$$\mu_{\widehat{\zeta}} \leq C\sqrt{\log q} \max_{1\leq j\leq r} \widehat{w}_{jj}^P.$$

By setting $u^2 = -2\log\alpha \max_{1\leq j\leq r} \widehat{w}_{jj}^P$, we have

$$\widehat{q}_{\mathcal{S}}(1-\alpha) \leq (C_0\sqrt{\log q} + \sqrt{-2\log\alpha}) \max_{1\leq j\leq r} \widehat{w}_{jj}^P,$$

Define the event G_ϵ as $G_\epsilon = \{\max_{1\leq j\leq r} \frac{|\widehat{w}_{jj}^P - w_{jj}^P|}{w_{jj}^P} \leq \epsilon\}$; as is shown in the proof of Theorem 2, $\max_{1\leq j\leq r} |\widehat{w}_{jj}^P - w_{jj}^P| = o(1)$, which indicates that $\max_{1\leq j\leq r} \frac{|\widehat{w}_{jj}^P - w_{jj}^P|}{w_{jj}^P} = o(1)$. For any $\epsilon > 0$, this event hold with probability tending to 1. Now under G_ϵ, it is obvious that

$$\widehat{q}_{\mathcal{S}}(1-\alpha) \leq (1+\epsilon)(C_0\sqrt{\log q} + \sqrt{-2\log\alpha}) \max_{1\leq j\leq r} w_{jj}^P, \tag{A.57}$$

Without loss of generality, we assume that $\frac{\sum_{t=1}^k \xi_{t12}\sqrt{n_t p}(\rho_{t12}-c_{t12})}{\sqrt{k}} = \max_{(i,j)\in\mathcal{S}} |\frac{\sum_{t=1}^k \xi_{tij}\sqrt{n_t p}(\rho_{tij}-c_{tij})}{\sqrt{k}}|$, it immediately follows that

$$\frac{\sum_{t=1}^k \xi_{t12}\sqrt{n_t p}(\rho_{t12} - c_{t12})}{\sqrt{k}} \geq C\sqrt{\log q} \max_{1\leq j\leq r} (w_{jj}^P)^{1/2}, \tag{A.58}$$

Also, we have

$$\begin{aligned}
&\mathbb{P}_{H_1}(\Psi_\alpha, G_\epsilon) \\
&= \mathbb{P}_{H_1}(\max_{(i,j)\in\mathcal{S}} |\frac{\sum_{t=1}^k \xi_{tij}\sqrt{n_t p}(\widehat{\rho}_{tij} - c_{tij})}{\sqrt{k}}| > \widehat{q}_{\mathcal{S}}(1-\alpha), G_\epsilon) \\
&\geq \mathbb{P}_{H_1}(\frac{\sum_{t=1}^k \xi_{t12}\sqrt{n_t p}(\widehat{\rho}_{t12} - c_{t12})}{\sqrt{k}} > \widehat{q}_{\mathcal{S}}(1-\alpha), G_\epsilon) \\
&= \mathbb{P}_{H_1}(G_\epsilon) - \mathbb{P}_{H_1}(\frac{\sum_{t=1}^k \xi_{t12}\sqrt{n_t p}(\widehat{\rho}_{t12} - c_{t12})}{\sqrt{k}} \leq \widehat{q}_{\mathcal{S}}(1-\alpha), G_\epsilon) \\
&= \mathbb{P}_{H_1}(G_\epsilon) - \mathbb{P}_{H_1}(\frac{\sum_{t=1}^k \xi_{t12}\sqrt{n_t p}(\widehat{\rho}_{t12} - \rho_{t12})}{\sqrt{k}} \leq \widehat{q}_{\mathcal{S}}(1-\alpha) - \frac{\sum_{t=1}^k \xi_{t12}\sqrt{n_t p}(\rho_{t12} - c_{t12})}{\sqrt{k}}, G_\epsilon),
\end{aligned}$$

By Equation (A.57), Equation (A.58) and noticing that C is a large enough constant and ϵ can be arbitrarily small, we have

$$\mathbb{P}_{H_1}(\Psi_\alpha, G_\epsilon) \geq \mathbb{P}_{H_1}(G_\epsilon) - \mathbb{P}_{H_1}(\frac{\sum_{t=1}^k \xi_{t12}\sqrt{n_t p}(\widehat{\rho}_{t12} - \rho_{t12})}{\sqrt{k}} \leq -C'\sqrt{\log q} \max_{1\leq j\leq r} (w_{jj}^P)^{1/2}, G_\epsilon),$$

for some constant C'. Since the first term tends to 1 while the second term tends to 0, we have

$$\mathbb{P}_{H_1}(\Psi_\alpha) \geq \mathbb{P}_{H_1}(\Psi_\alpha, G_\epsilon) \to 1,$$

Therefore we complete the proof. □

Proof of Corollary 1

Proof. The proof follows similarly as in Proof of Theorem 3. Noticing that $\text{sign}(\rho_{1ij}) = \ldots = \text{sign}(\rho_{dij})$ and $\text{sign}(c_{1ij}) = \ldots = \text{sign}(c_{dij})$, thus we omit ξ_{tij} here. For the first claim, simply notice that under the null, for any (i, j), we have

$$\mathbb{P}_{H_0'}\left(\frac{\sum_{t=1}^k \sqrt{n_t p}(|\widehat{\rho}_{tij}| - c)}{\sqrt{k}} > \widehat{q}_{\mathcal{S}}(1-\alpha)\right) \leq \mathbb{P}_{H_0'}\left(\frac{\sum_{t=1}^k \sqrt{n_t p}(|\widehat{\rho}_{tij} - \rho_{tij}|)}{\sqrt{k}} > \widehat{q}_{\mathcal{S}}(1-\alpha)\right)$$

Therefore we have

$$\begin{aligned}\mathbb{P}_{H_0'}(\Psi_\alpha') &= \mathbb{P}_{H_0'}\left(\max_{(i,j)\in\mathcal{S}} \frac{\sum_{t=1}^k \sqrt{n_t p}(|\widehat{\rho}_{tij}| - c)}{\sqrt{k}} > \widehat{q}_{\mathcal{S}}(1-\alpha)\right) \\ &\leq \mathbb{P}_{H_0'}\left(\max_{(i,j)\in\mathcal{S}} \frac{\sum_{t=1}^k \sqrt{n_t p}(|\widehat{\rho}_{tij} - \rho_{tij}|)}{\sqrt{k}} > \widehat{q}_{\mathcal{S}}(1-\alpha)\right) \to \alpha.\end{aligned}$$

For the second claim, we follow the proof in Theorem 3 to Equation (A.57). Without loss of generality we assume $\frac{\sum_{t=1}^k \sqrt{n_t p}(|\rho_{t12}|-c)}{\sqrt{k}} = \max_{(i,j)\in\mathcal{S}} \frac{\sum_{t=1}^k \sqrt{n_t p}(|\rho_{tij}|-c)}{\sqrt{k}}$. Likewise, we have

$$\frac{\sum_{t=1}^k \sqrt{n_t p}(|\rho_{t12}| - c)}{\sqrt{k}} \geq C\sqrt{\log q} \max_{1\leq j\leq r} (w_{jj}^P)^{1/2}, \tag{A.59}$$

Notice that,

$$\begin{aligned}\mathbb{P}_{H_1'}(\Psi_\alpha', G_\epsilon) &= \mathbb{P}_{H_1}\left(\max_{(i,j)\in\mathcal{S}} \frac{\sum_{t=1}^k \sqrt{n_t p}(|\widehat{\rho}_{tij}| - c)}{\sqrt{k}} > \widehat{q}_{\mathcal{S}}(1-\alpha), G_\epsilon\right) \\ &\geq \mathbb{P}_{H_1'}\left(\frac{\sum_{t=1}^k \sqrt{n_t p}(|\widehat{\rho}_{t12}| - c)}{\sqrt{k}} > \widehat{q}_{\mathcal{S}}(1-\alpha), G_\epsilon\right) \\ &= \mathbb{P}_{H_1'}\left(\frac{\sum_{t=1}^k \sqrt{n_t p}(|\widehat{\rho}_{t12}| - |\rho_{t12}|)}{\sqrt{k}} > -\frac{\sum_{t=1}^k \sqrt{n_t p}(|\rho_{t12}| - c)}{\sqrt{k}} + \widehat{q}_{\mathcal{S}}(1-\alpha), G_\epsilon\right) \\ &\geq \mathbb{P}_{H_1'}\left(-\frac{\sum_{t=1}^k \sqrt{n_t p}(|\widehat{\rho}_{t12} - \rho_{t12}|)}{\sqrt{k}} > -\frac{\sum_{t=1}^k \sqrt{n_t p}(|\rho_{t12}| - c)}{\sqrt{k}} + \widehat{q}_{\mathcal{S}}(1-\alpha), G_\epsilon\right)\end{aligned}$$

By Equation (A.57), (A.59), and notice that C is a large enough constant, we have

$$\mathbb{P}_{H_1'}(\Psi_\alpha', G_\epsilon) \to 1,$$

Therefore $\mathbb{P}_{H_1'}(\Psi_\alpha') \geq \mathbb{P}_{H_1'}(\Psi_\alpha', G_\epsilon) \to 1$ and we complete the proof. □

A.5 Tuning-free Method for Single MGGM

In single-session data case, we omit the index t here for clarity. For each fixed $1 \leq i \leq q$, the self-tuned Dantzig-constraint is defined as follows

$$D(i) = \{(\beta, \sigma) : \beta \in \mathbb{R}^{p-1}, \sigma > 0, \|\frac{1}{np}\sum_{k=1}^n \sum_{l=1}^p (\boldsymbol{X}_{l,-i}^{(k)})'(X_{li}^{(k)} - \boldsymbol{X}_{l,-i}^{(k)}\boldsymbol{\beta}_i)\|_\infty < \lambda\sigma\},$$

where $\lambda = \sqrt{\frac{\log \max(q,np)}{np}}$. Then the Self-tuned Dantzig estimator $(\widehat{\beta}_i, \widehat{\sigma}_i)$ is any solution to the following optimization problem

$$\min_{(\beta,\sigma)\in D(i)} (\|\beta\|_1 + c\sigma), \tag{A.60}$$

where c is a predefined constant which we will discuss about later.

With the regression as in Equation (A.60), following [27], we introduce a few definitions which will be required for our proposition and proof. For a support set $J \subseteq \{1, .., q-1\}$, we define the cone of dominant coordinates as

$$C_J^{(\gamma)} := \{\boldsymbol{\Delta} \in \mathbb{R}^{q-1} : \|\boldsymbol{\Delta}_{J^c}\|_1 \leq (1+\gamma)\|\boldsymbol{\Delta}_J\|_1\}.$$

We define $l_q - J_0 - block$ sensitivity as

$$\kappa_{q,J_0,J}^{(\gamma)} := \inf_{\boldsymbol{\Delta}\in C_J^{(\gamma)}:\|\boldsymbol{\Delta}_{J_0}\|_q=1} \|\boldsymbol{\Psi}\boldsymbol{\Delta}\|_\infty,$$

with $\boldsymbol{\Psi} = \widehat{\boldsymbol{V}}_{-i}$, where $\widehat{\boldsymbol{V}}$ is the sample estimate for spatial covariance matrix. We define the restricted eigenvalue as

$$\kappa_{RE,J}^{(\gamma)} := \inf_{\boldsymbol{\Delta}\in\mathbb{R}^{q-1}\setminus\{0\}:\boldsymbol{\Delta}\in C_J^{(\gamma)}} \frac{|\boldsymbol{\Delta}'\boldsymbol{\Psi}\boldsymbol{\Delta}|}{\|\boldsymbol{\Delta}_J\|_2^2}.$$

Proposition 3. *Define $\Delta\boldsymbol{\beta}_i = \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i$; set $\lambda = \sqrt{\frac{\log q}{np}}$, $c = \frac{2\lambda}{\kappa_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}^{1/2}}$ in self-tuned Dantzig selector; we have $\widehat{\boldsymbol{\beta}}_i$ are consistent with the following properties: for any $\delta_1, \delta_2 > 0$, there exists constants C_1, C_2 such that*

$$\mathbb{P}(\max_{1\leq i\leq q}\|\Delta\boldsymbol{\beta}_i\|_1 \geq C_1 s\lambda) \leq q^{-\delta_1},$$

$$\mathbb{P}(\max_{1\leq i\leq q}\|\Delta\boldsymbol{\beta}_i\|_2 \geq C_2\sqrt{s}\lambda) \leq q^{-\delta_2}.$$

Remark: In practice, we introduce one way to pick c. According to Proposition 4.2 in [27], we introduce $\kappa_{1,0}^{(\gamma)} := \frac{1}{s}\min_{k=1,\ldots,q-1}\{\min_{\Delta_k=1,\|\boldsymbol{\Delta}\|_1\leq(2+\gamma)s}\|\boldsymbol{\Psi}\boldsymbol{\Delta}\|_\infty\}$, which is a lower bound of $\kappa_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}^{\gamma}$, and can be computed by solving linear programs based on the data only. If s is unknown, we may choose an upper bound based on Assumption 4.

Proof. The proof can be separated into algebraic and probabilistic parts. Let's focus on the algebraic part on certain events. In the end we will show that the probabilities of those events are close to 1. Define the event G_i as

$$\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\varepsilon_{li,k}\boldsymbol{X}_{l,-i}^{(k)}\|_\infty \leq \lambda\phi_i,$$

where $\lambda = \sqrt{\frac{\log q}{np}}$, ϕ_i is some constant $O(1)$. Denote the difference of our estimator and true coefficients as $\Delta\boldsymbol{\beta}_i = \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i$. Then we have

$$\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}_{l,-i}^{(k)})'\boldsymbol{X}_{l,-i}^{(k)}\Delta\boldsymbol{\beta}_i\|_\infty \le \|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}_{l,-i}^{(k)})'(\boldsymbol{X}_{l,i}^{(k)}-(\boldsymbol{X}_{l,-i}^{(k)})'\widehat{\boldsymbol{\beta}}_i)\|_\infty$$
$$+\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}\varepsilon_{li,k}\boldsymbol{X}_{l,-i}^{(k)}\|_\infty$$
$$\le(\widehat{\sigma}_i+\phi_i)\lambda\le\lambda(2\phi_i+(\widehat{\sigma}_i-\phi_i)).$$

Note on event G_i, we have

$$\|\widehat{\boldsymbol{\beta}}_i\|_1+c\widehat{\sigma}_i\le\|\boldsymbol{\beta}_i\|_1+c\phi_i, \tag{A.61}$$

which further implies that

$$\|\Delta\boldsymbol{\beta}_{i,J^c(\boldsymbol{\beta}_i)}\|_1\le\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1+c(\phi_i-\widehat{\sigma}_i). \tag{A.62}$$

Therefore, we have

$$\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}_{l,-i}^{(k)})'\boldsymbol{X}_{l,-i}^{(k)}\Delta\boldsymbol{\beta}_i\|_\infty\le\lambda(2\phi_i+\frac{1}{c}(\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1-\|\Delta\boldsymbol{\beta}_{i,J^c(\boldsymbol{\beta}_i)}\|_1). \tag{A.63}$$

For every $\gamma>0$, we have the following two cases:

- $c\phi_i>\|\gamma\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1$;
- $c\phi_i\le\|\gamma\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1$.

For the first case, it directly implies that $\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1<\frac{c}{\gamma}\phi_i$; for the second case, combing Equation A.62 we have

$$\|\Delta\boldsymbol{\beta}_{i,J^c(\boldsymbol{\beta}_i)}\|_1\le(1+\gamma)\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1. \tag{A.64}$$

Hence $\Delta\boldsymbol{\beta}_i\in C^\gamma_{J(\boldsymbol{\beta_i})}$. By the definition of $\kappa^\gamma_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}$, we claim that with high probability,

$$\kappa^\gamma_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}\ge\frac{C}{s}, \tag{A.65}$$

where C is a constant. To see this, Lemma 4.1 in [27] gives that $\kappa_{RE,J(\boldsymbol{\beta}_i)}\le(2+\gamma)s\kappa^\gamma_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}$, and we only need to show that $\kappa_{RE,J(\boldsymbol{\beta}_i)}$ is lower bounded by a constant. For $\boldsymbol{\Delta}\in C^\gamma_{J(\boldsymbol{\beta_i})}$,

$$\boldsymbol{\Delta}'\boldsymbol{\Psi}\boldsymbol{\Delta}=\boldsymbol{\Delta}'\boldsymbol{V}_{-i}\boldsymbol{\Delta}+\boldsymbol{\Delta}'(\boldsymbol{\Psi}-\boldsymbol{V}_{-i})\boldsymbol{\Delta}\ge\frac{1}{c_e}\|\boldsymbol{\Delta}\|_2^2-\|\boldsymbol{\Psi}-\boldsymbol{V}_{-i}\|_\infty\|\boldsymbol{\Delta}\|_1^2$$
$$\ge\frac{1}{c_e}\|\boldsymbol{\Delta}\|_2^2-\|\boldsymbol{\Psi}-\boldsymbol{V}_{-i}\|_\infty(2+\gamma)^2s\|\boldsymbol{\Delta}\|_2^2.$$

By Lemma 2 and Assumption 4, $\|\boldsymbol{\Psi}-\boldsymbol{V}\|_\infty(2+\gamma)^2s\le\frac{1}{2c_e}$ with high probability. Therefore $\kappa_{RE,J(\boldsymbol{\beta}_i)}$ is on constant level indeed, i.e,

$$\frac{|\boldsymbol{\Delta}'\boldsymbol{\Psi}\boldsymbol{\Delta}|}{\|\boldsymbol{\Delta}_{J(\boldsymbol{\beta}_i)}\|_2^2}\ge\frac{|\boldsymbol{\Delta}'\boldsymbol{\Psi}\boldsymbol{\Delta}|}{\|\boldsymbol{\Delta}\|_2^2}\ge\frac{1}{2c_e}.$$

Next, continuing our calculation, Equation A.63 implies that

$$\kappa^{\gamma}_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1 \leq \lambda(2\phi_i + \frac{1}{c}\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1).$$

With the condition that $\kappa^{\gamma}_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)} > \frac{\lambda}{c}$,

$$\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1 \leq \frac{2\phi_i\lambda}{\kappa^{\gamma}_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)} - \frac{\lambda}{c}}.$$

Combining two cases we have

$$\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1 \leq \max\{\frac{c}{\gamma}\phi_i, \frac{2\phi_i\lambda}{\kappa^{\gamma}_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)} - \frac{\lambda}{c}}\}.$$

For simplicity, we pick $\gamma = 1/2$ and $c = \frac{2\lambda}{\kappa^{1/2}_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}}$. Then we have $\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1 \leq \frac{4\lambda}{\kappa^{1/2}_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}}\phi_i$ and according to Equation A.62,

$$\begin{aligned}\|\Delta\boldsymbol{\beta}_i\|_1 &= \|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1 + \|\Delta\boldsymbol{\beta}_{i,J^c(\boldsymbol{\beta}_i)}\|_1 \\ &\leq 2\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1 + c\phi_i \\ &\leq \frac{10\lambda}{\kappa^{1/2}_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}}\phi_i \leq \frac{10\phi_i}{C}\lambda s, \end{aligned} \tag{A.66}$$

where the last inequality follows from Equation A.65. Noticing that $\frac{1}{c}\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1 \leq 2\lambda\phi_i$, Equation A.63 becomes

$$\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}^{(k)}_{l,-i})'\boldsymbol{X}^{(k)}_{l,-i}\Delta\boldsymbol{\beta}_i\|_\infty \leq \lambda(2\phi_i + \frac{1}{c}\|\Delta\boldsymbol{\beta}_{i,J(\boldsymbol{\beta}_i)}\|_1) \leq 4\lambda\phi_i. \tag{A.67}$$

Therefore, combing the above equation with Equation A.66, we have the prediction error of Dantzig selector as

$$\begin{aligned}&\frac{1}{np}|\Delta\boldsymbol{\beta}_i'\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}^{(k)}_{l,-i})'\boldsymbol{X}^{(k)}_{l,-i}\Delta\boldsymbol{\beta}_i| \\ &\leq \|\Delta\boldsymbol{\beta}_i\|_1\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}^{(k)}_{l,-i})'\boldsymbol{X}^{(k)}_{l,-i}\Delta\boldsymbol{\beta}_i\|_\infty \\ &\leq \frac{40\lambda^2}{\kappa^{1/2}_{1,J(\boldsymbol{\beta}_i),J(\boldsymbol{\beta}_i)}}\phi_i^2. \end{aligned} \tag{A.68}$$

To proceed with $\|\Delta\boldsymbol{\beta}_i\|_2$, let us denote the event G_i' as

$$\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}^{(k)}_{l,-i})'\boldsymbol{X}^{(k)}_{l,-i} - \boldsymbol{V}_{-i}\|_\infty < \lambda\phi_i',$$

where $\boldsymbol{V}_{-i}$ is a submatrix of $\boldsymbol{V}$ with i-th row and column removed, and ϕ_i' is a constant $O(1)$. First, we notice that

$$\begin{aligned}
&\Delta\boldsymbol{\beta}_i' \frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}_{l,-i}^{(k)})'\boldsymbol{X}_{l,-i}^{(k)}\Delta\boldsymbol{\beta}_i \\
&= \Delta\boldsymbol{\beta}_i'(\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}_{l,-i}^{(k)})'\boldsymbol{X}_{l,-i}^{(k)} - \boldsymbol{V}_{-i})\Delta\boldsymbol{\beta}_i + \Delta\boldsymbol{\beta}_i'\boldsymbol{V}_{-i}\Delta\boldsymbol{\beta}_i \\
&\geq \frac{1}{c_e}\|\Delta\boldsymbol{\beta}_i\|_2^2 - \|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}_{l,-i}^{(k)})'\boldsymbol{X}_{l,-i}^{(k)} - \boldsymbol{V}_{-i}\|_\infty\|\Delta\boldsymbol{\beta}_i\|_1^2. \qquad \text{(A.69)}
\end{aligned}$$

Thus under G_i and G_i', we have

$$\begin{aligned}
\|\Delta\boldsymbol{\beta}_i\|_2^2 &\leq c_e\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}_{l,-i}^{(k)})'\boldsymbol{X}_{l,-i}^{(k)} - \boldsymbol{V}_{-i}\|_\infty\|\Delta\boldsymbol{\beta}_i\|_1^2 \\
&\quad + c_e\|\Delta\boldsymbol{\beta}_i\|_1\|\frac{1}{np}\sum_{k=1}^{n}\sum_{l=1}^{p}(\boldsymbol{X}_{l,-i}^{(k)})'\boldsymbol{X}_{l,-i}^{(k)}\Delta\boldsymbol{\beta}_i\|_\infty \\
&\leq c_e(\lambda\phi_i'\|\Delta\boldsymbol{\beta}_i\|_1^2 + 4\lambda\phi_i\|\Delta\boldsymbol{\beta}_i\|_1), \qquad \text{(A.70)}
\end{aligned}$$

where the last inequality follows from Equation A.67. Since Equation A.66 indicates that $\|\Delta\boldsymbol{\beta}_i\|_1$ is a $o(1)$ term, we know $\lambda\phi_i'\|\Delta\boldsymbol{\beta}_i\|_1^2 < \lambda\phi_i\|\Delta\boldsymbol{\beta}_i\|_1$. Therefore,

$$\|\Delta\boldsymbol{\beta}_i\|_2^2 \leq 5c_e\lambda\phi_i\|\Delta\boldsymbol{\beta}_i\|_1 \leq C's\lambda^2.$$

Therefore, following Equation A.66, we have

$$\|\Delta\boldsymbol{\beta}_i\|_2 \leq \sqrt{C'}\phi_i\sqrt{s}\lambda. \qquad \text{(A.71)}$$

As the probability of event G_i and G_i' is given in Lemma 2 and Lemma 3, from Assumption 4, Equation A.66 and Equation A.71, we conclude that for any $\delta_1, \delta_2 > 0$, there exists constants C_1, C_2 such that

$$\mathbb{P}(\max_{1\leq i\leq q}\|\Delta\boldsymbol{\beta}_i\|_1 \geq C_1 s\lambda) \leq q^{-\delta_1},$$

$$\mathbb{P}(\max_{1\leq i\leq q}\|\Delta\boldsymbol{\beta}_i\|_2 \geq C_2\sqrt{s}\lambda) \leq q^{-\delta_2}.$$

□

A.5.1 Evaluating Goodness of Fit

To assess the model fit, we calculate the mean log-likelihood of our model for each node along time averaged over all trials for the first session. From Fig. A.7(a), we observe that the top-left sub-region has the largest log-likelihood while the bottom-right sub-region has the lowest log-likelihood. This indicates that nodes with stronger connectivity play a more important role when fitting the model. Notice that there is another local minimum at the top-right (location

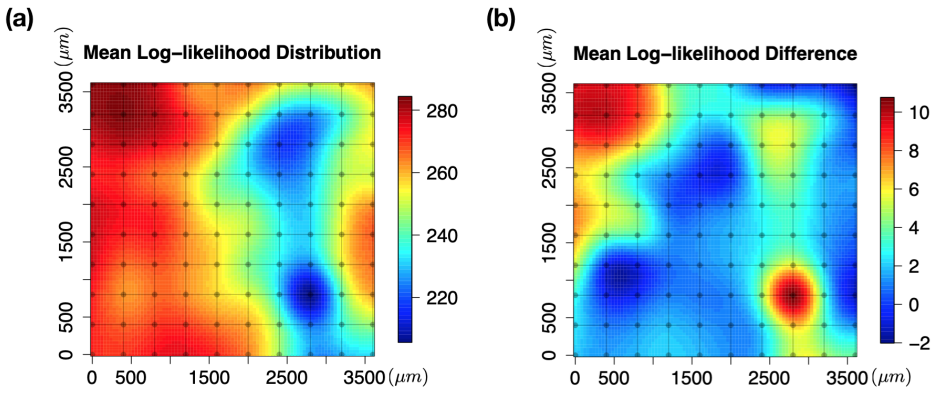

Figure A.7: Mean log-likelihood for each node along time averaged over all trials: left panel is the likelihood of the original fit, and right panel is the likelihood difference between the 3×3 *finer and original fit. Comparing two panels, log-likelihood values for sub-regions with largest log-likelihood and lowest log-likelihood are both greatly improved by adapting a local fit.*

$(2800, 3200)$) of the figure, which results from the fact that this area has a lower signal strength thus a lower weight when fitting the model parameters. Next, we divide the 10×10 array into nine parts: along x-axis and y-axis, we divide them into three subset (1,2,3), (4,5,6), (7,8,9,10), so that the 2D space consists of 3×3 sub-regions. For each sub-region, we fit a MGGM separately, and the mean log-likelihood difference between this fit and the original fit is in Fig. A.7(b). Most of the sub-regions correspond to positive values, which indicates that the fit improves; moreover, comparing two panels, log-likelihood values for sub-regions with the largest log-likelihood and the lowest log-likelihood are both greatly improved by this local fit model, which indicates that different factors may exist inside each sub-region and need to be addressed differently. This motivates us to apply a multi-factor model in the next Chapter so that each local mode can be effectively captured.

Appendix B

Appendix to Chapter 3

B.1 EM-algorithm to Fit LDFA-H

Initialization Let $\widehat{\boldsymbol{\theta}}^{(0)} = \{\widehat{\boldsymbol{\Sigma}}_1^{(0)}, \ldots, \widehat{\boldsymbol{\Sigma}}_q^{(0)}, \widehat{\boldsymbol{\Phi}}_{\mathcal{S}}^{1,(0)}, \widehat{\boldsymbol{\Phi}}_{\mathcal{S}}^{2,(0)}, \widehat{\boldsymbol{\Phi}}_{\mathcal{T}}^{1,(0)}, \widehat{\boldsymbol{\Phi}}_{\mathcal{T}}^{2,(0)}, \widehat{\boldsymbol{\beta}}^{1,(0)}, \widehat{\boldsymbol{\beta}}^{2,(0)}, \widehat{\boldsymbol{\mu}}^{1,(0)}, \widehat{\boldsymbol{\mu}}^{2,(0)}\}$ be the initial parameter value. Since the MPLE objective function for LDFA-H given in Eq. (3.9) is not guaranteed convex, an EM-algorithm may find a local minimum according to a choice of the initial value. Hence a good initialization is crucial to a successful estimation. Here we suggest an initialization by a canonical correlation analysis (CCA).

Let $\{\boldsymbol{X}^1[n], \boldsymbol{X}^2[n]\}_{n=1,\ldots,N}$ be N simultaneously recorded pairs of neural time series. We can view them as NT recorded pairs of multivariate random vectors $\{\boldsymbol{X}_{t,\cdot}^1[n], \boldsymbol{X}_{t,\cdot}^2[n]\}_{(n,t)\in[N]\times[T]}$. We obtain $\widehat{\boldsymbol{\beta}}_1^{1,(0)}$ and $\widehat{\boldsymbol{\beta}}_1^{2,(0)}$ by CCA as follows:

$$\widehat{\boldsymbol{\beta}}_1^{1,(0)}, \widehat{\boldsymbol{\beta}}_1^{2,(0)} = \underset{\widehat{\boldsymbol{\beta}}_1^1 \in \mathbb{R}^{p_1}, \beta_1^2 \in \mathbb{R}^{p_2}}{\operatorname{argmax}} \frac{\boldsymbol{\beta}_1^{1\top} \boldsymbol{S}^{12} \boldsymbol{\beta}_1^2}{\sqrt{\boldsymbol{\beta}_1^{1\top} \boldsymbol{S}^{11} \boldsymbol{\beta}_1^1} \sqrt{\boldsymbol{\beta}_1^{2\top} \boldsymbol{S}^{22} \boldsymbol{\beta}_1^2}} \tag{B.1}$$

where

$$\begin{aligned}
\boldsymbol{S}^{11} &= \frac{1}{NT} \sum_{n,t} (\boldsymbol{X}_{t,\cdot}^1[n] - \frac{1}{NT} \sum_{n,t} \boldsymbol{X}_{t,\cdot}^1[n])(\boldsymbol{X}_{t,\cdot}^1[n] - \frac{1}{NT} \sum_{n,t} \boldsymbol{X}_{t,\cdot}^1[n])^\top \\
\boldsymbol{S}^{22} &= \frac{1}{NT} \sum_{n,t} (\boldsymbol{X}_{t,\cdot}^2[n] - \frac{1}{NT} \sum_{n,t} \boldsymbol{X}_{t,\cdot}^2[n])(\boldsymbol{X}_{t,\cdot}^2[n] - \frac{1}{NT} \sum_{n,t} \boldsymbol{X}_{t,\cdot}^2[n])^\top \\
\boldsymbol{S}^{12} &= \frac{1}{NT} \sum_{n,t} (\boldsymbol{X}_{t,\cdot}^1[n] - \frac{1}{NT} \sum_{n,t} \boldsymbol{X}_{t,\cdot}^1[n])(\boldsymbol{X}_{t,\cdot}^2[n] - \frac{1}{NT} \sum_{n,t} \boldsymbol{X}_{t,\cdot}^2[n])^\top.
\end{aligned} \tag{B.2}$$

According to the equivalence between CCA and probablistic CCA shown by [7], it gives an estimate of the first latent factors

$$\widehat{\boldsymbol{Z}}_{\cdot,1}^{k,(0)}[n] = \widehat{\boldsymbol{\beta}}_1^{k,(0)} \boldsymbol{X}^k[n] \tag{B.3}$$

for $n = 1, \ldots, N$ and $k = 1, 2$. The initial second latent factors $\widehat{\boldsymbol{Z}}_2^{k,(0)}$ and the corresponding factor loading $\widehat{\boldsymbol{\beta}}_2^{k,(0)}$ is similarly set by the second pair of canonical variables, and so on. Then we

assign the empirical covariance matrix of $\{\widehat{\boldsymbol{Z}}_f^{1,(0)}[n], \widehat{\boldsymbol{Z}}_f^{2,(0)}[n]\}_{n\in[N]}$ to the initial latent covariance matrix $\widehat{\boldsymbol{\Sigma}}_f^{(0)}$ for $f = 1, \ldots, q$ and the matrix-variate normal estimate [67] on $\{\widehat{\boldsymbol{\epsilon}}^{k,(0)}[n] := \boldsymbol{X}^k[n] - \widehat{\boldsymbol{\beta}}^{k,(0)}\widehat{\boldsymbol{Z}}^{k,(0)}[n]\}_{n\in[N]}$ to $\widehat{\boldsymbol{\Phi}}_{\mathcal{T}}^{k,(0)}$ and $\widehat{\boldsymbol{\Phi}}_{\mathcal{S}}^{k,(0)}$ for $k = 1, 2$. Along $\widehat{\boldsymbol{\mu}}^{k,(0)} := \frac{1}{N}\sum_{n=1}^N \widehat{\boldsymbol{X}}^k[n]$, the above parameters comprises the initial parameter set $\widehat{\boldsymbol{\theta}}^{(0)}$.

However, we cannot run an E-step on the above parameter set because $\widehat{\boldsymbol{\Phi}}^{k,(0)}$ is not invertible. We instead pick one of its unidentifiable parameter sets $\widehat{\boldsymbol{\theta}}^{(0),\{\alpha^1,\alpha^2\}}$, defined in Eq. (3.8), with all $\widehat{\boldsymbol{\Phi}}^{k,(0)}$'s and $\widehat{\boldsymbol{\Sigma}}_f^{(0)}$'s invertible. Specifically, we take

$$\alpha_f^k = \frac{1}{2}\lambda_{\min}\left(\boldsymbol{\Sigma}_f^{1/2}\begin{bmatrix}\boldsymbol{\Phi}_{\mathcal{T}}^1 & 0\\ 0 & \boldsymbol{\Phi}_{\mathcal{T}}^2\end{bmatrix}^{-1}\boldsymbol{\Sigma}_f^{1/2}\right) \tag{B.4}$$

for $f = 1, \ldots, q$ and $k = 1, 2$ where $\lambda_{\min}(\boldsymbol{A})$ is the smallest eigenvalue of symmetric matrix $\boldsymbol{A}$. Henceforth, we notate $\widehat{\boldsymbol{\theta}}^{(0),\{\alpha^1,\alpha^2\}}$ by $\widehat{\boldsymbol{\theta}}^{(0)}$. For $t = 1, 2, \ldots$, we iterate the following E-step and M-step until convergence.

Another promising initialization is by finding time (t, s) on which the canonical correlation between $\boldsymbol{X}_{t,\cdot}^1$ and $\boldsymbol{X}_{s,\cdot}^2$ maximizes. i.e., we initialize $\widehat{\boldsymbol{\beta}}_1^{1,(0)}$ and $\widehat{\boldsymbol{\beta}}_1^{2,(0)}$ by

$$\widehat{\boldsymbol{\beta}}_1^{1,(0)}, \widehat{\boldsymbol{\beta}}_1^{2,(0)} = \underset{\boldsymbol{\beta}_1^1\in\mathbb{R}^{p_1}, \boldsymbol{\beta}_1^2\in\mathbb{R}^{p_2}}{\operatorname{argmax}} \frac{\boldsymbol{\beta}_1^{1\top}\boldsymbol{S}_{(t,s)}^{12}\boldsymbol{\beta}_1^2}{\sqrt{\boldsymbol{\beta}_1^{1\top}\boldsymbol{S}_{(t,t)}^{11}\boldsymbol{\beta}_1^1}\sqrt{\boldsymbol{\beta}_1^{2\top}\boldsymbol{S}_{(s,s)}^{22}\boldsymbol{\beta}_1^2}} \text{ such that } |t-s| < h_{\text{cross}}. \tag{B.5}$$

where

$$\begin{aligned}
\boldsymbol{S}_{(t,t)}^{11} &= \frac{1}{N}\sum_{n,t}(\boldsymbol{X}_{t,\cdot}^1[n] - \frac{1}{N}\sum_n \boldsymbol{X}_{t,\cdot}^1[n])(\boldsymbol{X}_{t,\cdot}^1[n] - \frac{1}{N}\sum_n \boldsymbol{X}_{t,\cdot}^1[n])^\top\\
\boldsymbol{S}_{(s,s)}^{22} &= \frac{1}{N}\sum_{n,s}(\boldsymbol{X}_{s,\cdot}^2[n] - \frac{1}{N}\sum_n \boldsymbol{X}_{t,\cdot}^2[n])(\boldsymbol{X}_{s,\cdot}^2[n] - \frac{1}{N}\sum_n \boldsymbol{X}_{s,\cdot}^2[n])^\top\\
\boldsymbol{S}_{(t,s)}^{12} &= \frac{1}{N}\sum_{n,t}(\boldsymbol{X}_{t,\cdot}^1[n] - \frac{1}{N}\sum_n \boldsymbol{X}_{t,\cdot}^1[n])(\boldsymbol{X}_{s,\cdot}^2[n] - \frac{1}{N}\sum_n \boldsymbol{X}_{s,\cdot}^2[n])^\top.
\end{aligned} \tag{B.6}$$

for $(t, s) \in [T] \times [T]$. Then the other parameters are initialized as above. We can even take an ensemble approach in which we fit LDFA-H on different initialized values and pick the estimate with the minimum cost function (Eq. (3.9)).

Now, for $r = 1, 2, \ldots$, we alternate an E-step and an M-step until the target parameter Π_f convergences.

E-step Given $\widehat{\boldsymbol{\theta}} := \widehat{\boldsymbol{\theta}}^{(r-1)}$ from the previous iteration, the conditional distribution of latent factors $\boldsymbol{Z}^1[n]$ and $\boldsymbol{Z}^2[n]$ with respect to observed data $\boldsymbol{X}^1[n]$ and $\boldsymbol{X}^2[n]$ on trial $n = 1, \ldots, N$ follows

$$\left(\boldsymbol{Z}_{\cdot,1}^1[n]; \boldsymbol{Z}_{\cdot,1}^2[n]; \ldots; \boldsymbol{Z}_{\cdot,q}^2[n]\right) \mid \boldsymbol{X}^1[n], \boldsymbol{X}^2[n] \sim \text{MVN}\left(\boldsymbol{m}_{\vec{Z}|X}^{(r)}[n], \boldsymbol{V}_{\vec{Z}|X}^{(r)}\right), \tag{B.7}$$

where

$$\boldsymbol{V}^{(r)}_{\vec{Z}|X} = \begin{pmatrix} \boldsymbol{V}^{(r)}_{Z_1,Z_1|X} & \cdots & \boldsymbol{V}^{(r)}_{Z_1,Z_q|X} \\ \vdots & \ddots & \vdots \\ \boldsymbol{V}^{(r)}_{Z_q,Z_1|X} & \cdots & \boldsymbol{V}^{(r)}_{Z_q,Z_q|X} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}^{(r)}_{Z_1,Z_1|X} & \cdots & \boldsymbol{W}^{(r)}_{Z_1,Z_q|X} \\ \vdots & \ddots & \vdots \\ \boldsymbol{W}^{(r)}_{Z_q,Z_1|X} & \cdots & \boldsymbol{W}^{(r)}_{Z_q,Z_q|X} \end{pmatrix}^{-1} \tag{B.8}$$

and

$$\begin{aligned} \boldsymbol{m}^{(r)}_{\vec{Z}|X}[n] &= \left(\boldsymbol{m}^{(r)}_{Z^1_1|X}; \boldsymbol{m}^{(r)}_{Z^1_2|X}; \ldots; \boldsymbol{m}^{(r)}_{Z^2_q|X}\right) \\ &= \boldsymbol{V}^{(r)}_{\vec{Z}|X} \left(\widehat{\boldsymbol{\beta}}^{1\top}_1 \widehat{\boldsymbol{\Gamma}}^1_{\mathcal{S}} \boldsymbol{X}^1[n] \widehat{\boldsymbol{\Gamma}}^1_{\mathcal{T}} \; ; \; \widehat{\boldsymbol{\beta}}^{2\top}_1 \widehat{\boldsymbol{\Gamma}}^2_{\mathcal{S}} \boldsymbol{X}^2[n] \widehat{\boldsymbol{\Gamma}}^2_{\mathcal{T}} \; ; \; \ldots \; ; \; \widehat{\boldsymbol{\beta}}^{2\top}_q \widehat{\boldsymbol{\Gamma}}^2_{\mathcal{S}} \boldsymbol{X}^2[n] \widehat{\boldsymbol{\Gamma}}^2_{\mathcal{T}}\right) \end{aligned} \tag{B.9}$$

given

$$\boldsymbol{W}^{(r)}_{Z_f,Z_g|X} = \begin{pmatrix} (\widehat{\boldsymbol{\beta}}^{1\top}_f \widehat{\boldsymbol{\Gamma}}^1_{\mathcal{S}} \widehat{\boldsymbol{\beta}}^1_g) \, \widehat{\boldsymbol{\Gamma}^1}_{\mathcal{T}} & 0 \\ 0 & (\widehat{\boldsymbol{\beta}}^{2\top}_f \widehat{\boldsymbol{\Gamma}}^2_{\mathcal{S}} \widehat{\boldsymbol{\beta}}^2_g) \, \widehat{\boldsymbol{\Gamma}}^2_{\mathcal{T}} \end{pmatrix} + \mathbb{I}_{\{f=g\}} \, \widehat{\boldsymbol{\Omega}}_f, \;\; \mathbb{I}_{\{f=g\}} = \begin{cases} 1, & f = g \\ 0, & \text{o.w.} \end{cases} \tag{B.10}$$

for $f, g = 1, \ldots, q$.

M-step We find $\widehat{\boldsymbol{\theta}}^{(r)}$ which maximize the conditional expectation of the penalized likelihood under the same constraints in Eq. (3.9), i.e.

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{(r)} = \operatorname{argmin} & \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{Z[n]|X[n],\widehat{\theta}^{(r-1)}} \left[\log p(\boldsymbol{X}^1[n], \boldsymbol{X}^2[n], \boldsymbol{Z}^1[n], \boldsymbol{Z}^2[n]; \widehat{\boldsymbol{\theta}}^{(r-1)})\right] \\ & + \sum_{f=1}^{q} \sum_{k,l=1}^{2} \left\|\Lambda^{kl}_f \odot \boldsymbol{\Pi}^{kl}_f\right\|_1 \;\; \text{s.t. } \widehat{\boldsymbol{\Gamma}}^k_{\mathcal{T}} \text{ is } (2h^k_\epsilon + 1)\text{-diagonal} \end{aligned} \tag{B.11}$$

where p is the probability density function of our model in Eqs. (3.1), (3.4) and (3.5) and the expectation $\mathbb{E}_{Z[n]|X[n],\widehat{\theta}^{(r-1)}}$ follows the conditional distribution in Eq. (B.7). Taking a block coordinate descent approach, we solve the optimization problem by alternating M1 - M4.

M1: With respect to latent precision matrices $\boldsymbol{\Omega}_f$, Eq. (B.11) reduces to a graphical Lasso problem,

$$\widehat{\boldsymbol{\Omega}}^{(r)}_f = \underset{\boldsymbol{\Omega}_f}{\operatorname{argmin}} \left\{ -\log\det(\boldsymbol{\Omega}_f) + \operatorname{tr}\left(\boldsymbol{\Omega}_f \left(\boldsymbol{V}^{(r)}_{Z_f|X} + \widehat{\mathbb{E}}[\boldsymbol{m}^{(r)}_{Z_f|X} \boldsymbol{m}^{(r)\top}_{Z_f|X}]\right)\right) + \sum_{k,l=1}^{2} \left\|\Lambda^{kl}_f \odot \Pi^{kl}_f\right\|_1 \right\} \tag{B.12}$$

for each $f = 1, \ldots, q$ where $\widehat{\mathbb{E}}[\boldsymbol{m}^{(r)}_{Z_f|X} \boldsymbol{m}^{(r)\top}_{Z_f|X}] = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{m}^{(r)}_{Z_f|X}[n] \; \boldsymbol{m}^{(r)\top}_{Z_f|X}[n]$. The graphical Lasso problem is solved by the P-GLASSO algorithm by [42].

M2: With respect to $\boldsymbol{\Gamma}^k$, Eq. (B.11) reduces to an estimation of matrix-variate normal model [67]. The estimation problem can be formulated as

$$\widehat{\boldsymbol{\Gamma}}^{k(r)}_{\mathcal{S}} = \frac{1}{T} \left(\widehat{\mathbb{E}} \left[\boldsymbol{m}^{(r)}_{\epsilon^k|X} \boldsymbol{m}^{(r)\top}_{\epsilon^k|X}\right] + \sum_{f,g=1}^{q} \operatorname{tr}(\boldsymbol{V}^{(r)}_{Z^k_f,Z^k_g|X}) \boldsymbol{\beta}^k_f \boldsymbol{\beta}^{k\top}_g \right) \tag{B.13}$$

and

$$\widehat{\mathbf{\Gamma}}_{\mathcal{T}}^{k(r)} = \underset{\mathbf{\Gamma}_{\mathcal{T}}^k}{\operatorname{argmin}} \left\{ \begin{array}{l} -\log\det(\mathbf{\Gamma}_{\mathcal{T}}^k) \\ +\dfrac{1}{p_k}\operatorname{tr}\Big(\mathbf{\Gamma}_{\mathcal{T}}^k\Big(\displaystyle\sum_{f,g=1}^{q}(\boldsymbol{\beta}_f^{k\top}\mathbf{\Gamma}_{\mathcal{S}}^k\boldsymbol{\beta}_g^k)\,\boldsymbol{V}_{Z_f^k,Z_g^k|X}^{(r)} + \widehat{\mathbb{E}}\Big[\boldsymbol{m}_{\epsilon^k|X}^{(r)\top}\mathbf{\Gamma}_{\mathcal{S}}^k\boldsymbol{m}_{\epsilon^k|X}^{(r)}\Big]\Big)\Big) \end{array} \right\} \quad \text{(B.14)}$$
$$\text{s.t. } \widehat{\mathbf{\Gamma}}_{\mathcal{T}}^k \text{ is } (2h_\epsilon^k+1)\text{-diagonal}$$

for each $k = 1, 2$ where $\boldsymbol{m}_{\epsilon^k|X}^{(r)} = \boldsymbol{X}^k - \boldsymbol{\beta}^k\boldsymbol{m}_{Z^k|X}^{(r)} - \boldsymbol{\mu}^k$ and $\widehat{\mathbb{E}}[A]$ is the empirical mean of a random matrix A. The estimation of $\mathbf{\Gamma}_{\mathcal{T}}^k$ under the bandedness constraint is tractable with modified Cholesky factor decomposition approach with bandwidth h_ϵ^k using the procedure by [5].

M3: With respect to $\boldsymbol{\beta}^k$, Eq. (B.11) reduces to a quadratic program

$$\widehat{\boldsymbol{\beta}}^{k(r)} = \arg\max_{\boldsymbol{\beta}^k} \left\{ \begin{array}{l} \displaystyle\sum_{t,s}\Gamma_{\mathcal{T},(t,s)}^k \operatorname{tr}\Big(\boldsymbol{\beta}^{k\top}\mathbf{\Gamma}_{\mathcal{S}}^k\boldsymbol{\beta}_k\,(\boldsymbol{V}_{Z_{t,\cdot}^k,\boldsymbol{Z}_{s,\cdot}^k|X}^{(r)} + \widehat{\operatorname{Cov}}[\boldsymbol{m}_{Z_{t,\cdot}^k|X}^{(r)}, \boldsymbol{m}_{Z_{s,\cdot}^k|X}^{(r)}])\Big) \\ -2\displaystyle\sum_{t,s}\Gamma_{\mathcal{T},(t,s)}^k \operatorname{tr}\Big(\mathbf{\Gamma}_{\mathcal{S}}^k\boldsymbol{\beta}^k\widehat{\operatorname{Cov}}[\boldsymbol{X}_{t,\cdot}^k, \boldsymbol{m}_{Z_{s,\cdot}^k|X}^{(r)}]\Big) \end{array} \right\} \quad \text{(B.15)}$$

where $\Gamma_{T,(t,s)}^k$ is the (t,s) entry in $\mathbf{\Gamma}_{\mathcal{T}}^k$ and $\widehat{\operatorname{Cov}}(A,B)$ is the empirical covariance matrix between random vectors A and B. The analytic form of the solution is given by

$$\boldsymbol{\beta}^k = \left(\sum_{t,s}\mathbf{\Gamma}_{\mathcal{T},(t,s)}^k(\boldsymbol{V}_{Z_{t,\cdot}^k,\boldsymbol{Z}_{s,\cdot}^k|X}^{(r)} + \widehat{\operatorname{Cov}}[\boldsymbol{m}_{Z_{t,\cdot}^k|X}^{(r)}, \boldsymbol{m}_{Z_{s,\cdot}^k|X}^{(r)}])\right)^{-1}\left(\sum_{t,s}\mathbf{\Gamma}_{\mathcal{T},(t,s)}^k\widehat{\operatorname{Cov}}[\boldsymbol{m}_{Z_{s,\cdot}^k|X}^{(r)}, \boldsymbol{X}_{t,\cdot}^k]\right) \quad \text{(B.16)}$$

M4: With resepct to μ^k, it is straight-forward that Eq. (B.11) yields

$$\widehat{\boldsymbol{\mu}}^{k(r)} = \widehat{\mathbb{E}}\left[\boldsymbol{X}^k - \sum_{f=1}^{q}\boldsymbol{\beta}_f^k\boldsymbol{m}_{Z_f^k|X}^{(r)\top}\right].$$

B.2 Simulation Details

We simulated realistic data with known cross-region connectivity as follows. Simulating $q = 1$ pair of latent time-series Z^k from Equation (3.2), we introduced an exact ground-truth for the inverse cross-correlation matrix $\mathbf{\Pi}_1^{12}$ by setting:

$$\mathbf{\Pi}_1 = \begin{bmatrix}(\boldsymbol{P}_{1,0}^{11})^{-1} & 0 \\ 0 & (\boldsymbol{P}_{1,0}^{22})^{-1}\end{bmatrix} + \begin{bmatrix}\boldsymbol{D}^1 & \mathbf{\Pi}_1^{12} \\ \mathbf{\Pi}_1^{12\top} & \boldsymbol{D}^2\end{bmatrix} \quad \text{(B.17)}$$

where $\boldsymbol{D}^1$ and $\boldsymbol{D}^2$ are diagonal matrices with elements $\boldsymbol{D}_{(t,t)}^1 = \sum_s \mathbf{\Pi}_{1,(t,s)}^{12}$ and $\boldsymbol{D}_{(s,s)}^2 = \sum_t \mathbf{\Pi}_{1,(t,s)}^{12}$, which ensures that the matrix on the right hand side is positive definite. The matrix on the left hand side contains the auto-precision matrices of the two latent time series, with elements simulated from the squared exponential function:

$$\boldsymbol{P}_{1,0}^{kk} = \left[\exp\left(-c^k(t-s)^2\right)\right]_{t,s} + \lambda\boldsymbol{I}_T, \quad \text{(B.18)}$$

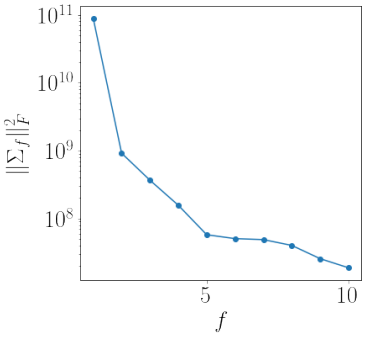

Figure B.1: Squared Frobenius norms of covariance matrix estimates, $\widehat{\Sigma}_f$, for all factors $f = 1, \ldots, 10$. Notice that the amplitudes of the top four factors dominate the others.

with $c^1 = 0.105$ and $c^2 = 0.142$, chosen to match the observed LFPs autocorrelations in the experimental dataset (Section 3.2.2). We added the regularizer λI_T, $\lambda = 1$, to render $\boldsymbol{P}^{kk}$ invertible. We designed the true inverse cross-correlation matrix Π^{12} to induce lead-lag relationship between $\boldsymbol{Z}^1$ and $\boldsymbol{Z}^2$ in two epochs as depicted in the right-most panel of Fig. 3.1(a). Specifically, the elements of $\boldsymbol{\Pi}^{12}$ were set:

$$\Pi^{12}_{(t,s)} = \begin{cases} -r, & \text{where } Z^1_{1,t} \text{ and } Z^2_{1,s} \text{ partially correlate,} \\ 0, & \text{elsewhere,} \end{cases} \tag{B.19}$$

where the association intensity $r = 0.6$ was chosen to match our cross-correlation estimate in the experimental data (Section 3.2.2). Finally, we rescaled $\boldsymbol{P}_1 = \boldsymbol{\Pi}_1^{-1}$ to have diagonal elements equal to one. The corresponding factor loading vector $\boldsymbol{\beta}_1^k$ was randomly generated from standard multivariate normal distribution and then scaled to have $\|\boldsymbol{\beta}_1^k\|_2 = 1$.

We generated the noise ϵ^k from the $N = 1000$ trials of the experimental data analyzed in Section 3.2.2. First, we permuted the trials in one region to remove cross-region correlations. Let $\{Y^1[n], Y^2[n]\}_{n=1,\ldots,N}$ be the permuted dataset. Then we contaminated the dataset with white noise to modulate the strength of noise correlation relative to cross-region correlations. i.e.

$$\boldsymbol{\epsilon}^k_{t,\cdot} = \boldsymbol{Y}^k_{t,\cdot} - \boldsymbol{\mu}^k_{t,\cdot} + \boldsymbol{\eta}^k_{t,\cdot}, \quad \boldsymbol{\eta}^k_{t,\cdot} \overset{\text{indep}}{\sim} \text{MVN}\left(0, \lambda_\epsilon \widehat{\text{Cov}}[\boldsymbol{Y}^k_{t,\cdot}]\right), \quad \text{and} \quad \boldsymbol{\mu}^k_{t,\cdot} = \widehat{\mathbb{E}}[\boldsymbol{Y}^k_{t,\cdot}] \tag{B.20}$$

where $\widehat{\mathbb{E}}[\boldsymbol{Y}^k_{t,\cdot}]$ and $\widehat{\text{Cov}}[\boldsymbol{Y}^k_{t,\cdot}]$ wer the empirical mean and covariance matrix of $Y^k_{t,\cdot}$, respectively, for $k = 1, 2, t = 1, \ldots, T$. The noise auto-correlation level was modulated by $\lambda_\epsilon \in \{2.78, 1.78, 0.44, 0.11\}$. We also obtained $\boldsymbol{\Sigma}_1$ by scaling P_1 so that $\boldsymbol{\Sigma}^{kk}_{1,(t,s)} = \boldsymbol{\beta}_1^{k\top} \boldsymbol{S}^k_t \boldsymbol{\beta}_1^k$. Putting all the pieces together, we generated observed time series by Eq. (3.1).

B.3 Experimental Data Analysis Details

We investigated the strength of each factor, which is characterized by $\boldsymbol{\Sigma}_f$, in Fig. B.1. Notice that the strength decreases fast initially and becomes relatively slower starting from the fifth factor. Therefore, we pick the top 4 factors in our result section.

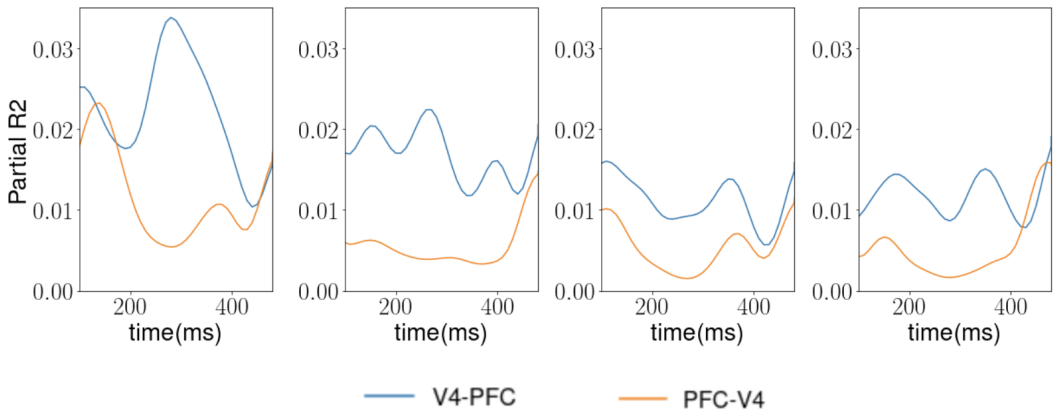

Figure B.2: Information flow by partial R^2 for the top four factors. In this figure, we characterize dynamic information flow in terms of partial R^2. We show dynamic information flow from $V4 \rightarrow PFC$ (blue) and $PFC \rightarrow V4$ (orange). In and out flows seem to peak at either the beginning or the end of the delay period, stronger $V4 \rightarrow PFC$ is identified, and different couplings of the two flows are also observed under this new definition. This figure echos with Fig. 3.4(b).

We also re-formulate the definition of information flow in the context of vector auto-regressive model. For the latent factor f in V4 at time t, consider the full regression model using the full history of latent variables in both area,

$$\boldsymbol{Z}^1_{t,f} \sim \boldsymbol{Z}^1_{1:t-1,f} + \boldsymbol{Z}^2_{1:t-1,f}$$

vs. the reduced model using history of latent variables in V4 only,

$$\boldsymbol{Z}^1_{t,f} \sim \boldsymbol{Z}^1_{1:t-1,f}.$$

The partial R^2 summarizes the contribution of PFC history to V4, thus can be viewed as information flow from V4 to PFC at time t. Dynamic information flow from V4 to PFC is defined similarly. The results are shown in Fig. B.2. Even from a different perspective, we reach to a similar conclusion as Fig. 3.4(b).