Journal of Machine Learning Research 21 (2020) 1-55 Submitted 6/18; Revised 12/19; Published 3/20

High-Dimensional Inference for Cluster-Based Graphical Models

Carson Eisenach EISENACH@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

Florentina Bunea FB238@CORNELL.EDU
Yang Ning YN265@CORNELL.EDU
Claudiu Dinicu CD535@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850, USA

Editor: Nicolas Vayatis

Abstract

Motivated by modern applications in which one constructs graphical models based on a very large number of features, this paper introduces a new class of cluster-based graphical models, in which variable clustering is applied as an initial step for reducing the dimension of the feature space. We employ model assisted clustering, in which the clusters contain features that are similar to the same unobserved latent variable. Two different cluster-based Gaussian graphical models are considered: the latent variable graph, corresponding to the graphical model associated with the unobserved latent variables, and the cluster-average graph, corresponding to the vector of features averaged over clusters. Our study reveals that likelihood based inference for the latent graph, not analyzed previously, is analytically intractable. Our main contribution is the development and analysis of alternative estimation and inference strategies, for the precision matrix of an unobservable latent vector Z. We replace the likelihood of the data by an appropriate class of empirical risk functions, that can be specialized to the latent graphical model and to the simpler, but under-analyzed, cluster-average graphical model. The estimators thus derived can be used for inference on the graph structure, for instance on edge strength or pattern recovery. Inference is based on the asymptotic limits of the entry-wise estimates of the precision matrices associated with the conditional independence graphs under consideration. While taking the uncertainty induced by the clustering step into account, we establish Berry-Esseen central limit theorems for the proposed estimators. It is noteworthy that, although the clusters are estimated adaptively from the data, the central limit theorems regarding the entries of the estimated graphs are proved under the same conditions one would use if the clusters were known in advance. As an illustration of the usage of these newly developed inferential tools, we show that they can be reliably used for recovery of the sparsity pattern of the graphs we study, under FDR control, which is verified via simulation studies and an fMRI data analysis. These experimental results confirm the theoretically established difference between the two graph structures. Furthermore, the data analysis suggests that the latent variable graph, corresponding to the unobserved cluster centers, can help provide more insight into the

understanding of the brain connectivity networks relative to the simpler, average-based, graph.

Keywords: Berry-Esseen bound, Graphical model, Latent variables, High-dimensional inference, Clustering, False discovery rate

1. Introduction

Over the last several decades, graphical models have become an increasingly popular method for understanding independence and conditional independence relationships between components of random vectors. More recently, the challenges posed by the estimation and statistical analysis of graphical models with many more nodes than the number of observations has led to renewed interest in these models, such as Meinshausen and Bühlmann (2006); Yuan and Lin (2007); Friedman et al. (2008); Verzelen (2008); Lam and Fan (2009); Rothman et al. (2008); Peng et al. (2009); Ravikumar et al. (2011); Yuan (2010); Cai et al. (2011); Sun and Zhang (2012); Liu et al. (2012); Xue and Zou (2012); Ning and Liu (2013); Cai et al. (2016); Tan et al. (2016); Fan et al. (2017); Yang et al. (2018); Feng and Ning (2019), to give only an incomplete list.

Nonetheless, when the dimension (number of nodes) grows very large and the sample size is small, the dependency among the components of a random vector may become weak, if it exists at all, and difficult to detect without additional information. If the dimension of the random vector is in the thousands, even if the dependency structure can be detected by an estimated graphical model, it can be difficult to interpret the results and extract meaningful scientific insights.

One solution to both of the aforementioned issues is to employ an initial dimension reduction procedure on the high dimensional vector. For example, in neuroscience applications, a typical functional magnetic resonance image (fMRI) consists of blood-oxygen-level-dependent (BOLD) activities measured at 200,000+ voxels of the brain, over a period of time. Instead of analyzing voxel-level data directly, scientists routinely cluster voxels into several regions of interest (ROI) with homogeneous functions using domain knowledge, and then carry out the analysis at the ROI-level. In this example, using the language of graphical models, the group structure of variables may boost the dependency signals. Similar pre-processing steps are used in other application domains, such as genomics, finance and economics.

Motivated by a rich set of applications, we consider variable clustering as the initial dimension reduction step applied to the observed vector $\boldsymbol{X} =: (X_1, \ldots, X_d) \in \mathbb{R}^d$. To the best of our knowledge, very little is known about the effect of clustering on downstream analysis and, consequently, on the induced graphical models. Our contribution is the provision of a framework that allows for such an analysis. We introduce cluster-based graphical models, show how they can be estimated and, furthermore, provide the asymptotic distribution of the edge strength estimates.

These models are built on the assumption that the observed variables $\boldsymbol{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ can be partitioned into K unknown clusters $G^* = \{G_1^*, \ldots, G_K^*\}$ such that variables in the same cluster share the same behavior. Following the intuition behind widely-used K-means type procedures, we define a population-level cluster as a group of variables that

are noise corrupted versions of a hub-like variable. This hub-like variable is not directly observable, and is treated as a latent factor.

Formally, we assume there exists a latent random vector $\boldsymbol{Z} \in \mathbb{R}^K$, with mean zero and covariance matrix $\mathrm{Cov}(\boldsymbol{Z}) = \mathbf{C}^*$, such that

$$\boldsymbol{X} = \mathbf{A}\boldsymbol{Z} + \boldsymbol{E}, \tag{1}$$

for a zero mean error vector $\boldsymbol{E}$ with independent entries. The entries of the $d \times K$ matrix $\mathbf{A}$ are given by $A_{jk} = \mathbb{I}\{j \in G_k^*\}$. A cluster of variables consist in those components of $\mathbf{X}$ with indices in the same G_k^*. We denote $\mathrm{Cov}(\boldsymbol{E}) = \mathbf{\Gamma}^*$, a diagonal matrix with entries $\Gamma_{jj}^* = \gamma_j^*$ for any $1 \leq j \leq d$. We also assume that the mean-zero noise $\boldsymbol{E}$ is independent of $\boldsymbol{Z}$. Bunea et al. (2018) show that the clusters are uniquely defined by the model in (1), provided that the smallest cluster contains at least two variables and $\mathbf{C}^*$ is strictly positive definite; this result holds irrespective of distributional assumptions.

To keep the presentation focused, in this work we assume that $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbf{C}^*)$ and $\boldsymbol{E} \sim \mathcal{N}(0, \mathbf{\Gamma}^*)$, which implies $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{\Sigma}^*)$ with $\mathbf{\Sigma}^* = \mathbf{A}\mathbf{C}^*\mathbf{A}^T + \mathbf{\Gamma}^*$. In this context we consider two related, but different, graphical models:

(i) The *latent variable graph*, associated with the sparsity pattern of the precision matrix

$$\mathbf{\Theta}^* := \mathbf{C}^{*-1} \tag{2}$$

of the Gaussian vector $\boldsymbol{Z} \in \mathbb{R}^K$. The latent variable graph encodes conditional independencies (CI's) among the unobserved, latent variables $\boldsymbol{Z}$.

(ii) The *cluster-average graph*, associated with the sparsity pattern of the precision matrix

$$\mathbf{\Omega}^* := \boldsymbol{S}^{*-1}, \tag{3}$$

where $\boldsymbol{S}^*$ is the covariance matrix of $\bar{\boldsymbol{X}} \in \mathbb{R}^K$, and $\bar{\boldsymbol{X}} =: (\bar{X}_1, \ldots, \bar{X}_K)$ is the within cluster average given by $\bar{X}_k =: \frac{1}{|G_k^*|}\sum_{i \in G_k^*} X_i$. The cluster-average graph encodes CI's among averages of observable random variables. In particular, we have

$$\boldsymbol{S}^* = \mathbf{C}^* + \bar{\mathbf{\Gamma}}^*,$$

where $\bar{\mathbf{\Gamma}}^* = \mathrm{diag}(\bar{\gamma}_1^*, ..., \bar{\gamma}_K^*)$ with $\bar{\gamma}_k^* = \frac{1}{|G_k^*|^2}\sum_{j \in G_k^*} \gamma_j^*$.

Although both these graphs correspond to vectors of dimension K and are constructed based on the partition G^*, they are in general different as the sparsity patterns of $\mathbf{\Theta}^*$ and $\mathbf{\Omega}^*$ will typically differ, and have different interpretations. Therefore it would be misleading to use one as a proxy for the other when drawing conclusions. For instance, in the neuroscience example, if we interpret each latent variable as the function of a ROI, then the latent variable graph encodes the CI relationships between functions, which is one question of scientific interest. The difference between $\mathbf{\Theta}^*$ and $\mathbf{\Omega}^*$ shows that this question will not be typically answered by studying the cluster-average graph, although that may be tempting to do by practitioners.

1.1. Our Contributions

Since the two cluster-based graphical models introduced above can both be of interest in a large array of applications, we provide inferential tools for both of them in this work. We assume that we observe n i.i.d. copies $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of $\boldsymbol{X}$. The focus of our work is on post-clustering and post-regularization inference for these two sparse precision matrices. To this end, we derive the asymptotic distribution of novel estimators of their entries. These estimators can be used to answer any inferential questions of interest at the edge strength level, or can be combined to provide sparsity pattern recovery under FDR control. In Section 4.3 we provide an instance of the latter.

Inference for the entries of a Gaussian precision matrix has received a large amount of attention in the past few years, most notably post-regularization inference, for instance Ren et al. (2013); Zhang and Zhang (2014); Jankova and van de Geer (2014); Gu et al. (2015); Barber and Kolar (2015); Janková and van de Geer (2017); Javanmard and Montanari (2013); van de Geer et al. (2013); Ning and Liu (2017); Cai and Guo (2017); Neykov et al. (2018); Ning et al. (2017); Fang et al. (2017). These works generalize the classical ideas of one-step estimation (Bickel, 1975) to the high-dimensional setup by first constructing a sparse estimator of the precision matrix via regularization, and then building de-sparsified updates that are asymptotically normal. The effect of the initial regularization step is controlled in the second step, and inference after regularization becomes valid.

In this paper we consider a similar estimation strategy, but differs from the existing literature in several important ways. In our work, we add another layer of data-dependent dimension reduction, via clustering, and provide a framework within which the variability induced by clustering can be controlled. Even after controlling for the clustering variability, we note that the existing procedures for estimation and, especially, post-regularization inference in Gaussian graphical models are not immediately applicable to our problem for the following reasons:

1. They are developed for variables that can be observed directly. From this perspective, they could, in principle, be applied to the cluster-average graph, but are not directly extendable to the latent graph;

2. To the best of our knowledge, all existing methods for precision matrix inference require the largest eigenvalue of the corresponding covariance matrix to be upper bounded by a constant. Such an assumption implies, in turn, that the Euclidean norm of each row of the covariance matrix is bounded, which reduces significantly the parameter space for which inference is valid. The assumption holds, for instance, when the number of variables is bounded, or when the entries of each row are appropriately small.

To overcome these limitations, we take a different approach in this work, that allows us to lift unpleasant technical conditions associated with other procedures, while maintaining the validity of inference for both the latent and the average graph. We summarize our main contributions below.

1. Methods for estimation tailored to high dimensional inference in latent, cluster-based, graphical models. We develop a new estimation strategy tailored to

our final goal, that of constructing approximately Gaussian estimators for the entries of the precision matrices $\mathbf{\Theta}^*$ and $\mathbf{\Omega}^*$ given in (2) and (3) above. Although we work under the assumption that the data is Gaussian, likelihood based estimators may be unsatisfactory, because their analysis can require stringent assumptions, as explained above (see also Jankova and van de Geer (2014)), or may become analytically intractable, as argued in Section 3.3, for the latent graph. We propose a method that mimics very closely the principles underlying the construction of an efficient score function for estimation in the presence of high dimensional nuisance parameters (see for instance Bickel et al. (1993)), but we do not base it on the corresponding likelihood-derived quantities. The underlying principles are explained in Section 3.1. To the best of our knowledge, this is the first estimation method of the latent precision matrix, that can be analyzed theoretically for inferential purposes. As an added benefit of our estimation framework, the same principles can be applied for the estimation, and analysis, of the cluster-average graph.

2. The analysis of estimators of the precision matrix of unobserved cluster centers: Berry-Esseen-type bounds for Gaussian approximations. We verify, theoretically, that the estimators constructed with an inferential goal in mind do indeed have the desired properties. To this end, we derive the asymptotic distribution of the proposed entry-wise estimates of the latent graph, and also of the cluster-average graph. Moreover, we quantify the speed with which this limit is obtained, which we show to be proportional to $1/\sqrt{n}$ in both cases. We do so by establishing Berry-Esseen type bounds on the difference between the cumulative distribution function of our estimators and that of a standard Gaussian random variable that are valid for each K, d and n, and are presented in Theorems 3 and 4, respectively. As immediate applications, we can construct approximate confidence intervals for one or a finite number of entries of the latent or average graph, or known linear functionals of such entries. While the answers we thus provide at the average graph level are similar to existing results, established for the full CI graph based on all d nodes, for instance by (Ren et al., 2013; Zhang and Zhang, 2014; Jankova and van de Geer, 2014; Gu et al., 2015; Barber and Kolar, 2015; Janková and van de Geer, 2017; Javanmard and Montanari, 2013; van de Geer et al., 2013; Ning and Liu, 2017; Cai and Guo, 2017; Neykov et al., 2018), the results for the latent graph are, to the best of our knowledge, the first such results in the literature.

We note, furthermore, that the average cluster graph can be viewed as a graph with observable nodes, the cluster averages, only after the clusters are estimated from the data. Our theoretical analysis takes this step into account. We discuss, in Section 2.2, clustering methods tailored to model (1), where the number of clusters K is unknown and is allowed to grow with n. Using the results of Bunea et al. (2018), these methods yield a partition $\widehat{G} = G^*$, with high probability, provided that $\lambda_{\min}(\mathbf{C}^*) > c$, for a small positive quantity c made precise in Section 2.2. A lower bound on the smallest eigenvalue of the covariance matrix is the minimal condition under which inference in *any* graphical model can be performed. Therefore, consistent clustering via the model (1) does not require a further reduction of the parameter space for which the more standard post-regularized inference can be developed. Moreover, as Section 4 shows, asymptotic inference based on the estimated clusters reduces to asymptotic inference relative to the true clusters, G^*, without any need for data splitting. This fact holds true for both the average and the latent variable graph, and is in sharp contrast with a phenomenon often encountered in post-model selection inference, such as

in variable selection in linear regression (Lockhart et al., 2014; Lee et al., 2013; Tibshirani et al., 2014). In that case, reducing inference to the consistently selected set of variables can only be justified over a reduced part of the parameter space (Bunea, 2004), and is therefore not a popular practice.

Another technical contribution is that the asymptotic normality of the estimators is established under relaxed conditions. Unlike the existing literature on de-biased inference on graphical models (Jankova and van de Geer, 2014; Janková and van de Geer, 2017), we do not require the bounded operator norm condition for the covariance matrix such as $\lambda_{\max}(\boldsymbol{S}^*) \leq C$ for cluster-average graph. As shown by Jankova and van de Geer (2014) and explained above, the analysis of the multivariate Gaussian likelihood may require stringent assumptions for cluster-average graph and becomes intractable for the latent graph. By using the proposed pseudo-likelihood function which has a much simpler form, we can remove the unpleasant assumption on the bounded operator norm. In addition, we re-analyze the CLIME estimator (Cai et al., 2011) $\widehat{\boldsymbol{\Omega}}_{\cdot k}$ (the kth column of $\widehat{\boldsymbol{\Omega}}$) under our Assumption 4.1 and 4.2, which is used as the initial estimator for inference. As explained in Section 3.2, we can show that the CLIME estimator satisfies $\|\widehat{\boldsymbol{\Omega}}_{\cdot k} - \boldsymbol{\Omega}^*_{\cdot k}\|_1 \lesssim s_1 \sqrt{\frac{\log(K \vee n)}{n}}$, where s_1 is the sparsity of $\boldsymbol{\Omega}^*_{\cdot k}$. This result does not require the bounded operator norm or matrix L_1 norm condition and can be of independent interest.

To illustrate how one can use these newly developed inferential tools, we focus on the estimation of the sparsity pattern of the graphs, which can be equivalently viewed as a multiple-testing problem. It is well known that the exact sparsity pattern can be recovered, with high probability, only if the entries of each precision matrix are above the minimax optimal noise level $O(\sqrt{\log d/n})$ (Ravikumar et al., 2011; Meinshausen and Bühlmann, 2006). Since our aim is inference on the sparsity pattern without further restrictions on the parameter space, the next best type of error that we can control is the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). In Section 4.3 we use these results for pattern recovery under FDR control, and explain the effect of the asymptotic approximations on this quantity.

This paper is organized as follows. Section 2 below contains a brief summary of existing results on model-assisted clustering, via model (1). Section 3 describes the estimation procedures for the latent variable graph and the cluster-average graph, respectively. In Section 4 we establish Berry-Esseen type central limit theorems for the estimators derived in Section 3, and provide bounds on the FDR associated with each graphical model under study, respectively. Section 5 gives numerical results using both simulated and real data sets.

2. Background

2.1. Notation

The following notation is adopted throughout this paper. Let d denote the ambient dimension, n the sample size, K the number of clusters and m the minimum cluster size. The matrix $\mathbf{C}^*$ denotes the population covariance of the latent vector $\boldsymbol{Z}$. Likewise, the matrices $\boldsymbol{\Gamma}^*$, $\boldsymbol{\Sigma}^*$, $\boldsymbol{\Theta}^*$, $\boldsymbol{S}^*$ and $\boldsymbol{\Omega}^*$ denote population-level quantities.

For $\mathbf{v} = (v_1, ..., v_d)^T \in \mathbb{R}^d$, and $1 \leq q \leq \infty$, we define $\|\mathbf{v}\|_q = (\sum_{i=1}^d |v_i|^q)^{1/q}$, $\|\mathbf{v}\|_0 = |\text{supp}(\mathbf{v})|$, where $\text{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$ and $|A|$ is the cardinality of a set A. Denote $\|\mathbf{v}\|_\infty = \max_{1\leq i\leq d} |v_i|$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$. Assume that $\mathbf{v}$ can be partitioned as $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$. Let $\nabla f(\mathbf{v})$ denote the gradient of the function $f(\mathbf{v})$, and $\nabla_1 f(\mathbf{v}) = \partial f(\mathbf{v})/\partial \mathbf{v}_1$. Similarly, let $\nabla^2 f(\mathbf{v})$ denote the Hessian of the function $f(\mathbf{v})$ and $\nabla^2_{12} f(\mathbf{v}) = \partial^2 f(\mathbf{v})/\partial \mathbf{v}_1 \partial \mathbf{v}_2$.

For a $d \times d$ matrix $\mathbf{M} = [M_{jk}]$, let $\|\mathbf{M}\|_{\max} = \max_{jk} |M_{jk}|$, $\|\mathbf{M}\|_1 = \sum_{jk} |M_{jk}|$, and $\|\mathbf{M}\|_\infty = \max_k \sum_j |M_{jk}|$. If the matrix $\mathbf{M}$ is symmetric, then $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ are the minimal and maximal eigenvalues of $\mathbf{M}$. Let $[d] = \{1, 2, ..., d\}$. For any $j \in [d]$, we denote the jth row and jth column of $\mathbf{M}$ as $\mathbf{M}_{j\cdot}$ and $\mathbf{M}_{\cdot j}$, respectively. Similarly, let $\mathbf{M}_{-j,-k}$ be the sub-matrix of $\mathbf{M}$ with the j^{th} row and k^{th} column removed. The notation $\mathcal{S}^{d\times d}$ refers to the set of all real, symmetric $d \times d$ matrices. Likewise, $\mathcal{S}^{d\times d}_+ \subset \mathcal{S}^{d\times d}$ is the positive semi-definite cone. We use $\otimes$ and $\circ$ to denote the Kronecker and Hadamard product of two matrices, respectively; we also may write $\mathbf{M}^{\otimes 2} = \mathbf{M} \otimes \mathbf{M}$. Let $\mathbf{e}_j$ denote the vector of all zeros except for a one in the j^{th} position. The vector $\mathbf{1}$ is the vector of all ones. $a \vee b = \max(a, b)$.

2.2. Model Assisted Variable Clustering

In this section we review existing results on variable clustering. Bunea et al. (2018) showed that if we use model (1) to define clusters of variables, these clusters are uniquely defined, up to label switching, so long as $m =: \min_{1\leq k\leq K} |G^*_k| \geq 2$ and the components of the latent vector $\boldsymbol{Z}$ are different almost surely, or equivalently

$$\Delta(\mathbf{C}^*) =: \min_{j<k} \mathbb{E}(Z_j - Z_k)^2 > 0.$$

Since

$$\Delta(\mathbf{C}^*) = \min_{j<k} (\mathbf{e}_j - \mathbf{e}_k)^T \mathbf{C}^* (\mathbf{e}_j - \mathbf{e}_k) \geq 2\lambda_{\min}(\mathbf{C}^*),$$

the clusters are uniquely defined as soon as $\lambda_{\min}(\mathbf{C}^*) > 0$, which is the minimal condition under which one can study properties of the corresponding precision matrix.

In addition, Bunea et al. (2018) developed two algorithms, PECOK and COD, that are shown to recover the clusters exactly, from n i.i.d. copies $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of $\boldsymbol{X}$, as soon as

$$\lambda_{\min}(\mathbf{C}^*) \geq c,$$

for a positive quantity c that approaches 0 as n grows. For the COD procedure,

$$c = O\left(\|\mathbf{\Sigma}^*\|_{\max} \sqrt{\log(d \vee n)/n}\right).$$

On the other hand, for the PECOK procedure

$$c = O\left(\|\mathbf{\Gamma}^*\|_{\max} \sqrt{K \log(d \vee n)/mn}\right),$$

which can be much smaller when one has a few, balanced, clusters.

These values of c are shown to be minimax or near-minimax optimal for cluster recovery. We refer to Theorems 3 and 4 in Bunea et al. (2018) for the precise expressions and details. Under these minimal conditions on $\lambda_{\min}(\mathbf{C}^*)$, the exact recovery of the clusters holds with

probability larger than $1 - 1/(d \vee n)$. Because these conditions are sufficiently weak, we are able to show, in Section 4, that inference in cluster-based graphical models is not hampered by the clustering step.

For completeness, we outline the PECOK algorithm below, which consists in a convex relaxation of the K-means algorithm, further tailored to estimation of clusters $G^* = \{G_1^*, \ldots, G_K^*\}$ defined via the interpretable model (1). The PECOK algorithm consists in the following three steps:

1. Compute an estimator $\widetilde{\mathbf{\Gamma}}$ of the matrix $\mathbf{\Gamma}^*$.

2. Solve the semi-definite program (SDP)

$$\widehat{\mathbf{B}} = \underset{\mathbf{B} \in \mathcal{D}}{\operatorname{argmax}} \langle \widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Gamma}}, \mathbf{B} \rangle, \tag{4}$$

   where $\widehat{\mathbf{\Sigma}}$ is the sample covariance matrix and

$$\mathcal{D} := \left\{ \mathbf{B} \in R^{d \times d} : \begin{array}{l} \bullet\ \boldsymbol{B} \succcurlyeq 0 \ \text{ (symmetric and positive semidefinite)} \\ \bullet\ \sum_a B_{ab} = 1, \ \forall b \\ \bullet\ B_{ab} \geq 0, \ \forall a, b \\ \bullet\ \operatorname{tr}(\mathbf{B}) = K \end{array} \right\}. \tag{5}$$

3. Compute $\widehat{G}$ by applying a clustering algorithm on the rows (or equivalently columns) of $\widehat{\mathbf{B}}$.

The construction of an accurate estimator $\widetilde{\mathbf{\Gamma}}$ of $\mathbf{\Gamma}^*$, before the cluster structure is known, is a crucial step for guaranteeing the statistical optimality of the PECOK estimator. Its construction is given in Bunea et al. (2018), and included in Appendix F, for the convenience of the reader.

We will employ an efficient algorithm for solving (4). Standard black-box SDP solvers, for a fixed precision, exhibit $\mathcal{O}(d^7)$ running time on (4), which is prohibitively expensive. Eisenach and Liu (2019) recently introduced the FORCE algorithm, which requires worst case $\mathcal{O}(d^6 K^{-2})$ time to solve the SDP, and in practice often performs the clustering rapidly.

The key idea behind the FORCE algorithm is that an optimal solution to (4) can be attained by first transforming (4) into an eigenvalue problem, and then using a first-order method. Iterations of the first-order method are interleaved with a dual step to round the current iterate to an integer solution of the clustering problem, and then searches for an optimality certificate. By using knowledge of both the primal and the dual SDPs, FORCE is able to find the solution much faster than a standard SDP solver. We refer to Eisenach and Liu (2019) for the detailed algorithm.

3. Estimation of Cluster-based Graphical Models

In this section, we propose a unified estimation approach, that utilizes similar loss functions for estimation and inference in the cluster-average and the latent variable graphs. We first describe our general principle, and then demonstrate its application to the two graphical models.

3.1. One-step Estimators for High-Dimensional Inference

Assume that we observe n i.i.d. realizations $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ of $\boldsymbol{X} \in \mathbb{R}^d$. Let $Q(\boldsymbol{\beta}, \boldsymbol{X})$ denote a known mapping of $\boldsymbol{\beta}$ and $\boldsymbol{X}$ to $\mathbb{R}$, where $\boldsymbol{\beta}$ is a q-dimensional unknown parameter of the distribution of $\boldsymbol{X}$. Often this Q is referred to as the loss function. We define the target parameter $\boldsymbol{\beta}^*$ as

$$\boldsymbol{\beta}^* = \operatorname{argmin} \mathbb{E}(Q(\boldsymbol{\beta}, \boldsymbol{X})).$$

Next, let us partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$, where $\theta \in \mathbb{R}$ is the univariate parameter of interest, and $\boldsymbol{\gamma} \in \mathbb{R}^{q-1}$ is a nuisance parameter. Our goal is to construct a $n^{1/2}$-consistent and asymptotically normal estimator for θ in high-dimensional models with $q = \dim(\boldsymbol{\beta}) \gg n$. In this case, the dimension of the nuisance parameter γ is large, which makes the inference on θ challenging. We start from the empirical risk function over n observations defined as

$$Q_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} Q(\boldsymbol{\beta}, \boldsymbol{X}_i). \tag{6}$$

One standard choice for Q_n is the negative log-likelihood function of the data. In this work, we conduct inference based on an alternative loss function, as the analysis of the log-likelihood can require unpleasant technical conditions that we would like to avoid, as discussed in Sections 4.1. That said, we mimic likelihood principles as much as possible, in order to make intuitive the construction below. For these reasons we will refer to $Q_n(\boldsymbol{\beta})$ as the negative *pseudo-likelihood* function.

For now we leave $Q_n(\boldsymbol{\beta})$ unspecified – a detailed discussion of its selection for inference in the latent variable graph and the cluster-average graph will be given in the following two subsections. In terms of Q_n, we define the pseudo-information matrix for one observation as $\mathbf{I}(\boldsymbol{\beta}) = \mathbb{E}(\nabla^2 Q(\boldsymbol{\beta}^*, \boldsymbol{X}_i))$. We can partition this matrix as

$$\mathbf{I}(\boldsymbol{\beta}) = \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix}, \tag{7}$$

with the partitions corresponding to those of $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$.

When Q_n is the negative log-likelihood function, and the dimension of the parameter is independent of n, then $h(\boldsymbol{\theta}; \boldsymbol{\gamma})$ given by (8) is called the *efficient score function* for θ, and classical theory shows that it admits solutions that are consistent, asymptotically normal and attain the information bound given by the reciprocal of (9) (Van der Vaart, 1998; Bickel et al., 1993).

With these goals in mind, we similarly define the corresponding pseudo-score function for estimating θ in the presence of the nuisance parameter $\boldsymbol{\gamma}$ as

$$\begin{aligned} h(\theta; \boldsymbol{\gamma}) &= \nabla_1 Q_n(\boldsymbol{\beta}) - \mathbf{I}_{12}\mathbf{I}_{22}^{-1} \nabla_2 Q_n(\boldsymbol{\beta}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \Big(\nabla_1 Q(\boldsymbol{\beta}, \boldsymbol{X}_i) - \mathbf{I}_{12}\mathbf{I}_{22}^{-1} \nabla_2 Q(\boldsymbol{\beta}, \boldsymbol{X}_i) \Big) \end{aligned} \tag{8}$$

and define the pseudo information of θ, in the presence of the nuisance parameter $\boldsymbol{\gamma}$, as

$$\mathbf{I}_{1|2} = \mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{I}_{21}. \tag{9}$$

When the dimension of $\boldsymbol{\gamma}$ is fixed, one can easily estimate $\mathbf{I}_{12}$ and $\mathbf{I}_{22}$ in (8) by their sample versions $\widehat{\mathbf{I}}_{12}$ and $\widehat{\mathbf{I}}_{22}$. However, such simple procedure fails when the dimension of $\boldsymbol{\gamma}$ is greater than the sample size, as $\widehat{\mathbf{I}}_{22}$ is rank deficient. To overcome this difficulty, rather than estimating $\mathbf{I}_{12}$ and $\mathbf{I}_{22}^{-1}$ separately, we directly estimate

$$\mathbf{w}^T = \mathbf{I}_{12}\mathbf{I}_{22}^{-1} \tag{10}$$

by

$$\widehat{\mathbf{w}} = \operatorname{argmin} \|\mathbf{w}\|_1, \quad \text{s.t.} \quad \|\nabla_{12}^2 Q_n(\widehat{\boldsymbol{\beta}}) - \mathbf{w}^T \nabla_{22}^2 Q_n(\widehat{\boldsymbol{\beta}})\|_\infty \leq \lambda', \tag{11}$$

where λ' is a non-negative tuning parameter, and $\widehat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\boldsymbol{\gamma}})$ is an initial estimator, which is usually defined case by case, for a given model. Then, we can plug $\widehat{\mathbf{w}}$ and $\widehat{\boldsymbol{\gamma}}$ into the pseudo-score function, which gives

$$\widehat{h}(\theta, \widehat{\boldsymbol{\gamma}}) = \nabla_1 Q_n(\theta, \widehat{\boldsymbol{\gamma}}) - \widehat{\mathbf{w}}^T \nabla_2 Q_n(\theta, \widehat{\boldsymbol{\gamma}}). \tag{12}$$

Following the Z-estimation principle (Van der Vaart, 1998; Bickel et al., 1993), one could define the final estimator of θ as the solution of the pseudo-score function $\widehat{h}(\theta, \widehat{\boldsymbol{\gamma}})$. However, in many examples, the pseudo-score function $\widehat{h}(\theta, \widehat{\boldsymbol{\gamma}})$ may have multiple solutions and it becomes unclear which root serves as a consistent estimator; see Small et al. (2000) for further discussion of the general estimating function context. To bypass this issue, we consider the following simple one-step estimation approach. Given the initial estimator $\widehat{\theta}$ from the partition of $\widehat{\boldsymbol{\beta}}$, we perform a Newton-Raphson update based on the pseudo-score function $\widehat{h}(\theta, \widehat{\boldsymbol{\gamma}})$, to obtain $\widetilde{\theta}$, which is traditionally referred to as a one-step estimator by Bickel (1975). Specifically, we construct

$$\widetilde{\theta} = \widehat{\theta} - \widehat{\mathbf{I}}_{1|2}^{-1} \widehat{h}(\widehat{\theta}, \widehat{\boldsymbol{\gamma}}), \tag{13}$$

where $\widehat{\mathbf{I}}_{1|2}$ is an estimator of the partial information matrix $\mathbf{I}_{1|2}$. In Sections 3.2 and 3.3 below we show that, under appropriate conditions, the one-step estimator $\widetilde{\theta}$ constructed using the empirical risk functions Q_n – defined in (16) and (31), respectively – satisfies

$$n^{1/2}(\widetilde{\theta} - \theta^*) = -\mathbf{I}_{1|2}^{-1} n^{1/2} h(\boldsymbol{\beta}^*) + o_p(1). \tag{14}$$

By applying the central limit theorem to $h(\boldsymbol{\beta}^*)$, we establish the asymptotic normality of $\widetilde{\theta}$ in Theorems 3 and 4.

When $Q_n(\boldsymbol{\beta})$ is the negative log-likelihood of the data, Ning and Liu (2017) successfully used this approach and the resulting estimator $\widetilde{\theta}$ is asymptotically equivalent to the de-biased estimator in Zhang and Zhang (2014); van de Geer et al. (2013). As we explain in the following subsections, analysis based on the log-likelihood becomes intractable for the latent graphical model and requires stringent technical conditions for the cluster-average graphical model. To overcome this difficulty, we employ pseudo-score functions derived from (16) and (31). The resulting one-step estimator still attains the information bound established in the literature, and more importantly requires weaker technical assumptions than the existing methods. In addition to (14), we derive explicitly the speed at which the normal approximation is attained.

3.2. Estimation of the Cluster-Average Graph

Recall that we assume $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbf{C}^*)$ and $\boldsymbol{E} \sim \mathcal{N}(0, \boldsymbol{\Gamma}^*)$, which implies $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^*)$ with $\boldsymbol{\Sigma}^* = \mathbf{A}\mathbf{C}^*\mathbf{A}^T + \boldsymbol{\Gamma}^*$. The within-cluster average $\bar{\boldsymbol{X}} =: (\bar{X}_1, \ldots, \bar{X}_K) \in \mathbb{R}^K$ is given by $\bar{X}_k =: \frac{1}{|G_k^*|}\sum_{i \in G_k^*} X_i$, corresponding to the population level clusters. Because $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma}^*)$, we can verify that $\bar{\boldsymbol{X}} \sim \mathcal{N}(0, \boldsymbol{S}^*)$, where

$$\boldsymbol{S}^* = \mathbf{C}^* + \bar{\boldsymbol{\Gamma}}^*, \tag{15}$$

and $\bar{\boldsymbol{\Gamma}}^* = \mathrm{diag}(\bar{\gamma}_1^*, ..., \bar{\gamma}_K^*)$ with $\bar{\gamma}_k^* = \frac{1}{|G_k^*|^2}\sum_{j \in G_k^*} \gamma_j^*$. Recall that the precision matrix of $\bar{\boldsymbol{X}}$ is

$$\boldsymbol{\Omega}^* = \boldsymbol{S}^{*-1} = (\mathbf{C}^* + \bar{\boldsymbol{\Gamma}}^*)^{-1}.$$

In this section we give the construction of the estimators of the cluster-average graph corresponding to $\bar{\boldsymbol{X}}$. Specifically, we use the generic strategy outlined in the previous section in order to construct $n^{1/2}$-consistent and asymptotically normal estimators for each component $\boldsymbol{\Omega}_{t,k}^*$ of the precision matrix $\boldsymbol{\Omega}^*$, for $1 \le t < k \le K$. For the estimation of each entry, the remaining $K(K+1)/2 - 1$ parameters in $\boldsymbol{\Omega}^*$ are treated as nuisance parameters.

Since we observe n i.i.d. samples of $\boldsymbol{X} \in \mathbb{R}^p$, if the clusters and their number were known, then we would implicitly observe n i.i.d. samples of $\bar{\boldsymbol{X}} \in \mathbb{R}^K$. A priori, the clusters are not known else this problem would simply reduce to the standard setting. However to explain our method, we first assume that clustering is given, and then show how to lift this assumption.

Following our general principle, we would naturally tend to choose the negative log-likelihood function of the cluster-averages $(\bar{\boldsymbol{X}}_1, ..., \bar{\boldsymbol{X}}_n)$ as the empirical risk function $Q_n(\boldsymbol{\beta})$ in (6). Along this line, Jankova and van de Geer (2014) proposed the de-biased estimator for Gaussian graphical models. However, the inference requires the irrepresentable condition (Ravikumar et al., 2011) on $\boldsymbol{S}^*$, which can be restrictive. The alternative methods proposed by Ren et al. (2013); Janková and van de Geer (2017) imposed the condition that the largest eigenvalue of $\boldsymbol{S}^*$ is bounded. These technical conditions on $\boldsymbol{S}^*$ are difficult to justify and can be avoided by using our approach. We propose to estimate each sparse row of $\boldsymbol{\Omega}^*$ as explained below.

Let $\bar{\boldsymbol{S}} = n^{-1}\sum_{i=1}^n \bar{\boldsymbol{X}}_i \bar{\boldsymbol{X}}_i^T$ denote the sample covariance matrix of $\bar{\boldsymbol{X}}_i$. When K is small, the maximum likelihood estimator of $\boldsymbol{\Omega}^*$ is $\bar{\boldsymbol{S}}^{-1}$, which can be viewed as the solution of the following equation $\bar{\boldsymbol{S}}\boldsymbol{\Omega} - \mathbf{I}_K = 0$. Thus, in the low dimensional setting, this equation defines the maximum likelihood estimator. Since we are only interested in $\boldsymbol{\Omega}_{t,k}^*$, we can extract the kth column from the left hand side of the above equation, and use it as the pseudo-score function $\boldsymbol{U}_n(\boldsymbol{\Omega}_{\cdot k}) = \bar{\boldsymbol{S}}\boldsymbol{\Omega}_{\cdot k} - \mathbf{e}_k$. To apply the inference strategy in Section 3.1, we need to construct a valid empirical risk function $Q_n(\boldsymbol{\Omega}_{\cdot k})$ such that $\nabla Q_n(\boldsymbol{\Omega}_{\cdot k}) = \boldsymbol{U}_n(\boldsymbol{\Omega}_{\cdot k})$.

Simple algebra shows that a possible choice is

$$Q_n(\boldsymbol{\Omega}_{\cdot k}) = \frac{1}{2}\boldsymbol{\Omega}_{\cdot k}^T \bar{\boldsymbol{S}} \boldsymbol{\Omega}_{\cdot k} - \mathbf{e}_k^T \boldsymbol{\Omega}_{\cdot k} = \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{2}\boldsymbol{\Omega}_{\cdot k}^T \bar{\boldsymbol{X}}_i \bar{\boldsymbol{X}}_i^T \boldsymbol{\Omega}_{\cdot k} - \mathbf{e}_k^T \boldsymbol{\Omega}_{\cdot k}\right), \tag{16}$$

which we view in the sequel as the empirical risk corresponding to the population level risk

$$\mathbb{E}Q(\boldsymbol{\Omega}_{\cdot k}, \bar{\boldsymbol{X}}) = \frac{1}{2}\boldsymbol{\Omega}_{\cdot k}^T \boldsymbol{S}^* \boldsymbol{\Omega}_{\cdot k} - \mathbf{e}_k^T \boldsymbol{\Omega}_{\cdot k}, \tag{17}$$

based on the loss function

$$Q(\mathbf{\Omega}_{\cdot k}, \bar{\boldsymbol{X}}) =: \frac{1}{2}\mathbf{\Omega}_{\cdot k}^T \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T \mathbf{\Omega}_{\cdot k} - \mathbf{e}_k^T \mathbf{\Omega}_{\cdot k}. \tag{18}$$

Since

$$\nabla \mathbb{E} Q(\mathbf{\Omega}^*_{\cdot k}, \bar{\boldsymbol{X}}) = \boldsymbol{S}^* \mathbf{\Omega}^*_{\cdot k} - \mathbf{e}_k = 0 \tag{19}$$

and

$$\nabla^2 \mathbb{E} Q(\mathbf{\Omega}^*_{\cdot k}, \bar{\boldsymbol{X}}) = \boldsymbol{S}^*, \tag{20}$$

then the population risk $\mathbb{E}Q(\mathbf{\Omega}^*_{\cdot k}, \bar{\boldsymbol{X}})$ has the rows $\mathbf{\Omega}^*_{\cdot k}$ of the target precision matrix $\mathbf{\Omega}^*$ as the unique minimizers, as desired, provided that $\boldsymbol{S}^*$ is positive definite, an assumption we make in Section 4.1.

We note that the choice of the empirical risk $Q_n(\cdot)$ and that of the corresponding pseudo-score $\boldsymbol{U}_n(\cdot)$ is not unique. We chose the particular form (16) because it is quadratic in $\mathbf{\Omega}_{\cdot k}$, which greatly simplifies the theoretical analysis and leads to weaker technical assumptions. Moreover, the property (19) is the same as that of the score function corresponding to the negative log-likelihood function, supporting our terminology.

We use the general strategy presented in Section 3.1 to construct estimators that employ the empirical risk $Q_n(\cdot)$ defined by (16) above. We first recall that $Q_n(\cdot)$ depends on the unknown cluster structure G^* via $\bar{\boldsymbol{X}}_i$. We note that in general the estimated group $\widehat{G}_k$ may differ from G^*_k by a label permutation. For notational simplicity, we ignore this label permutation issue and treat $\widehat{G}_k$ as an estimate of G^*_k (rather than G^*_j for some $j \neq k$). To define our estimator of $\mathbf{\Omega}^*_{t,k}$, we first replace $\bar{\boldsymbol{X}}_i$ by $\widehat{\boldsymbol{X}}_i$ and denote $\widehat{\boldsymbol{S}} = n^{-1}\sum_{i=1}^n \widehat{\boldsymbol{X}}_i \widehat{\boldsymbol{X}}_i^T$, where $\widehat{X}_{ik} = \frac{1}{|\widehat{G}_k|}\sum_{j\in\widehat{G}_k} X_{ij}$.

Let (t,k) be arbitrary, fixed. Replacing $\bar{\boldsymbol{S}}$ by $\widehat{\boldsymbol{S}}$ in $Q_n(\cdot)$, we follow Section 3.1 to define the pseudo-score function

$$h(\mathbf{\Omega}_{\cdot k}) = \mathbf{v}_t^{*T}(\widehat{\boldsymbol{S}}\mathbf{\Omega}_{\cdot k} - \mathbf{e}_k), \tag{21}$$

where $\mathbf{v}_t^*$ is a K-dimensional vector with $(\mathbf{v}_t^*)_t = 1$ and $(\mathbf{v}_t^*)_{-t} = -\mathbf{w}_t^*$ with $\mathbf{w}_t^* = (\boldsymbol{S}^*_{-t,-t})^{-1}\boldsymbol{S}^*_{-t,t}$, which is consistent with the definition in (10) above. To make inferences based on $h(\mathbf{\Omega}_{\cdot k})$, we further need to estimate $\mathbf{w}_t^*$ and $\mathbf{\Omega}^*_{\cdot k}$. Following (11), an estimate of $\mathbf{w}_t^*$ is given by

$$\widehat{\mathbf{w}}_t = \operatorname{argmin}\|\mathbf{w}\|_1, \quad \text{s.t} \quad \|\widehat{\boldsymbol{S}}_{t,-t} - \mathbf{w}^T\widehat{\boldsymbol{S}}_{-t,-t}\|_\infty \leq \lambda', \tag{22}$$

where λ' is a tuning parameter. Then we can define $\widehat{\mathbf{v}}_t$ accordingly, and

$$\widehat{h}(\mathbf{\Omega}_{\cdot k}) = \widehat{\mathbf{v}}_t^T(\widehat{\boldsymbol{S}}\mathbf{\Omega}_{\cdot k} - \mathbf{e}_k). \tag{23}$$

Recall that the construction of the one-step estimator (13) requires an initial estimator of $\mathbf{\Omega}^*_{\cdot k}$.

To be concrete, we consider the following initial estimator of $\mathbf{\Omega}^*_{\cdot k}$,

$$\widehat{\mathbf{\Omega}}_{\cdot k} = \operatorname{argmin}\|\boldsymbol{\beta}\|_1, \quad \text{s.t} \quad \|\widehat{\boldsymbol{S}}\boldsymbol{\beta} - \mathbf{e}_k\|_{\max} \leq \lambda, \tag{24}$$

where λ is a tuning parameter. This estimator has the same form as the CLIME estimator for the k-th column of $\mathbf{\Omega}$ (Cai et al., 2011). However, unlike the CLIME estimator which

requires $\lambda \asymp \|\mathbf{\Omega}^*_{\cdot k}\|_1\sqrt{\log K/n}$, in our Theorem 3 we assume $\lambda = C\sqrt{\log(K \vee n)/n}$, where C only depends on the minimum eigenvalue of $\mathbf{C}^*$ and the largest diagonal entries of $\mathbf{C}^*$ and $\mathbf{\Gamma}^*$ which are assumed bounded by constants in Assumptions 1 and 2. With this choice of λ, we show in Lemma 15 in Appendix A that

$$\|\widehat{\mathbf{\Omega}}_{\cdot k} - \mathbf{\Omega}^*_{\cdot k}\|_1 \lesssim s_1\sqrt{\frac{\log(K \vee n)}{n}}, \tag{25}$$

with high probability, where s_1 is the sparsity level of $\mathbf{\Omega}^*_{\cdot k}$. The sharp concentration of the gradient and Hessian of the empirical risk function provided in Lemma 14 is the key for this result. As a comparison, Theorem 6 in Cai et al. (2011) only implies $\|\widehat{\mathbf{\Omega}}_{\cdot k} - \mathbf{\Omega}^*_{\cdot k}\|_1 \lesssim s_1\|\mathbf{\Omega}^*_{\cdot k}\|_1^2\sqrt{\frac{\log K}{n}}$. For many sparse matrices, the ℓ_1 norm of a column, $\|\mathbf{\Omega}^*_{\cdot k}\|_1$, can grow to infinity with K or s_1, and thus (61) gives a faster rate.

In this case, and when $\lambda_{\min}(C^*) > c$, Lemma 14 is instrumental in showing that the extra $\|\mathbf{\Omega}^*_{\cdot k}\|_1^2$ factor in the rate of the original CLIME estimator can be avoided, whereas if only the marginal components of $\boldsymbol{X}$ are assumed to be sub-Gaussian, as in Cai et al. (2011), it may be unavoidable, without further conditions on $\mathbf{\Omega}^*$. We direct the reader to Appendix G for a more detailed discussion of the distinction between our results and those in Cai et al. (2011, 2016).

Leveraging the block matrix inverse formula, we can show that the partial pseudo information matrix reduces to $\mathbf{I}_{1|2} = 1/\Omega^*_{t,t}$. Finally, the one-step estimator is defined as

$$\widetilde{\Omega}_{t,k} = \widehat{\Omega}_{t,k} - \widehat{h}(\widehat{\mathbf{\Omega}}_{\cdot k})\widehat{\Omega}_{t,t}, \tag{26}$$

in accordance with (13). In Section 4, we show that under mild regularity conditions $n^{1/2}(\widetilde{\Omega}_{t,k} - \Omega^*_{t,k}) \rightsquigarrow N(0, s^2_{tk})$, where $s^2_{tk} = \Omega^{*2}_{t,k} + \Omega^*_{t,t}\Omega^*_{k,k}$. If $\widehat{s}^2_{tk} = \widehat{\Omega}^2_{t,k} + \widehat{\Omega}_{t,t}\widehat{\Omega}_{k,k}$ is a consistent estimator of the asymptotic variance, then a $(1-\alpha)\times 100\%$ confidence interval for $\Omega_{t,k}$ is

$$[\widetilde{\Omega}_{t,k} - z_{1-\alpha/2}\widehat{s}_{tk}/n^{1/2}, \widetilde{\Omega}_{t,k} + z_{1-\alpha/2}\widehat{s}_{tk}/n^{1/2}],$$

where z_α is the α-quantile of a standard normal distribution.

Equivalently, we can use the scaled test statistics $\widetilde{\Omega}_{t,k}$ to construct a test for $H_0 : \Omega^*_{t,k} = 0$ versus $H_1 : \Omega^*_{t,k} \neq 0$ with α significance level. Namely, the null hypothesis is rejected if and only if the above $(1-\alpha)\times 100\%$ confidence interval does not contain 0. We will employ such tests in Section 4.

3.3. Latent Variable Graph

Recall that the structure of the latent variable graph is encoded by the sparsity pattern of $\mathbf{\Theta}^* = \mathbf{C}^{*-1}$, which is generally different from the cluster-average group as $\mathbf{C}^{*-1}$ and $\mathbf{\Omega}^* = (\mathbf{C}^* + \bar{\mathbf{\Gamma}}^*)^{-1}$ may have different sparsity patterns. In this section, we focus on the inference on the component $\Theta^*_{t,k}$, for some $1 \le t < k \le K$. Similar to the cluster-average graph, we first discuss the likelihood approach. The negative log-likelihood corresponding to model (1) indexed by the parameter $(\mathbf{\Theta}, \mathbf{\Gamma})$ is, up to some additive and multiplicative constants,

$$\ell(\mathbf{\Theta}, \mathbf{\Gamma}) = \log|(\mathbf{A}\mathbf{\Theta}^{-1}\mathbf{A}^T + \mathbf{\Gamma})| + \operatorname{tr}(\widehat{\mathbf{\Sigma}}(\mathbf{A}\mathbf{\Theta}^{-1}\mathbf{A}^T + \mathbf{\Gamma})^{-1}),$$

where $\widehat{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$. It is straightforward to show that the Fisher information matrix for $(\mathbf{\Theta}, \mathbf{\Gamma})$ is given by

$$\mathbf{I}(\mathbf{\Theta}, \mathbf{\Gamma}) = \begin{bmatrix} (\mathbf{M}^*\mathbf{A}^T\mathbf{\Gamma}^{*-1}\mathbf{\Sigma}^*\mathbf{\Gamma}^{*-1}\mathbf{A}\mathbf{M}^*)^{\otimes 2} & (\mathbf{M}^*\mathbf{A}^T\mathbf{\Gamma}^{*-1}\mathbf{\Sigma}^{-1}\mathbf{F}^{*T})^{\otimes 2}\mathbf{D}_d \\ \mathbf{D}_d^T(\mathbf{F}^*\mathbf{\Sigma}^{*-1}\mathbf{\Gamma}^{*-1}\mathbf{A}\mathbf{M}^*)^{\otimes 2} & \mathbf{D}_d^T(\mathbf{F}^*\mathbf{\Sigma}^{*-1}\mathbf{F}^{*T})^{\otimes 2}\mathbf{D}_d \end{bmatrix}, \tag{27}$$

where $\mathbf{D}_d = (\mathbf{I}_d \otimes 1_d^T) \circ (1_d^T \otimes \mathbf{I}_d)$, $\mathbf{M}^* = (\mathbf{\Theta}^* + \mathbf{A}^T\mathbf{\Gamma}^{*-1}\mathbf{A})^{-1}$ and $\mathbf{F}^* = \mathbf{I}_d - \mathbf{A}\mathbf{M}^*\mathbf{A}^T\mathbf{\Gamma}^{*-1}$. As seen in Section 3.1, the inference based on the likelihood or equivalently efficient score function (8) requires the estimation of $\mathbf{I}_{12}\mathbf{I}_{22}^{-1}$ which, given the complicated structure of the information matrix (27), becomes analytically intractable.

A solution to this problem is inference based on an empirical risk function similar to (16), but tailored to the latent variable graph. With a slight abuse of notation, and reasoning as in (19) and (20), we notice that, for each k,

$$\mathbb{E}Q(\mathbf{\Theta}_{\cdot k}, \boldsymbol{X}) = \frac{1}{2}\mathbf{\Theta}_{\cdot k}^T\mathbf{C}^*\mathbf{\Theta}_{\cdot k} - \mathbf{e}_k^T\mathbf{\Theta}_{\cdot k}, \tag{28}$$

has the target $\mathbf{\Theta}^*_{\cdot k}$ as a unique minimizer, where the loss function $Q(\mathbf{\Theta}_{\cdot k}, \boldsymbol{X})$ is defined as

$$Q(\mathbf{\Theta}_{\cdot k}, \boldsymbol{X}) = \frac{1}{2}\mathbf{\Theta}_{\cdot k}^T\bar{\mathbf{C}}\mathbf{\Theta}_{\cdot k} - \mathbf{e}_k^T\mathbf{\Theta}_{\cdot k}, \tag{29}$$

and the matrix $\bar{\mathbf{C}} := (\bar{C}_{jk})_{j,k}$ has entries

$$\bar{C}_{jk} = \frac{1}{|G_j^*||G_k^*|}\sum_{a\in G_j^*, b\in G_k^*}(X_aX_b - \bar{\Gamma}_{ab}), \tag{30}$$

and $\bar{\Gamma}_{ab} = 0$ if $a \neq b$ and $\bar{\Gamma}_{aa} = X_aX_a - \frac{1}{|G_k^*|-1}\sum_{a\in G_k^*, a\neq j} X_aX_j$. Since $\mathbb{E}(\bar{\mathbf{C}}) = \mathbf{C}^*$, the risk relative to the loss function in (29) is indeed (28), and the empirical risk is

$$Q_n(\mathbf{\Theta}_{\cdot k}) = \frac{1}{n}\sum_{i=1}^n(\frac{1}{2}\mathbf{\Theta}_{\cdot k}^T\bar{\mathbf{C}}^{(i)}\mathbf{\Theta}_{\cdot k} - \mathbf{e}_k^T\mathbf{\Theta}_{\cdot k}), \tag{31}$$

where $\bar{\mathbf{C}}^{(i)}$ is obtained by replacing $\boldsymbol{X}$ in $\bar{C}_{jk}$ by $\boldsymbol{X}_i$. Similar to the cluster-average graph, $Q_n(\mathbf{\Theta}_{\cdot k})$ also depends on the unknown cluster structure. We estimate G_k^* by $\widehat{G}_k$, and define $\widehat{\mathbf{\Gamma}} = (\widehat{\Gamma}_{ab})$, where $\widehat{\Gamma}_{ab} = 0$ if $a \neq b$ and $\widehat{\Gamma}_{aa} = \frac{1}{n}\sum_{i=1}^n(X_{ia}X_{ia} - \frac{1}{|\widehat{G}_k|-1}\sum_{a\in\widehat{G}_k, a\neq j} X_{ia}X_{ij})$, and $\widehat{\mathbf{C}} = (\widehat{C}_{jk})$ where $\widehat{C}_{jk} = \frac{1}{n}\sum_{i=1}^n(\frac{1}{|\widehat{G}_j||\widehat{G}_k|}\sum_{a\in\widehat{G}_j, b\in\widehat{G}_k}(X_{ia}X_{ib} - \widehat{\Gamma}_{ab}))$. We replace $\bar{\mathbf{C}}$ by $\widehat{\mathbf{C}}$ in (31) above and follow exactly the strategy of Section 3.2, with $\widehat{\boldsymbol{S}}$ replaced by $\widehat{\mathbf{C}}$, to construct the corresponding pseudo-score function $\widehat{h}(\mathbf{\Theta}_{\cdot k})$, similarly to (23), and the initial estimator $\widehat{\mathbf{\Theta}}_{\cdot k}$, similarly to (24). We combine these quantities, following the general strategy (13), as above, to obtain the final one-step estimator of $\Theta^*_{t,k}$, defined as

$$\widetilde{\Theta}_{t,k} = \widehat{\Theta}_{t,k} - \widehat{h}(\widehat{\mathbf{\Theta}}_{\cdot k})\widehat{\Theta}_{t,t}, \tag{32}$$

after observing that, in this case, $\mathbf{I}_{1|2} = 1/\mathbf{\Theta}^*_{t,t}$.

Although the form of this estimator is similar to (26), derived for the cluster-average graph, the study of the asymptotic normality of $\widetilde{\Theta}_{t,k}$ reveals that its asymptotic variance is much more involved, as will be discussed in detail in Section 4.2.2.

4. Main Theoretical Results

4.1. Assumptions

In this section we state the two assumptions under which all our results are proved.

Assumption 1 *The covariance matrix* $\mathbf{C}^*$ *of* $\boldsymbol{Z}$ *satisfies:* $c_1 \leq \lambda_{\min}(\mathbf{C}^*)$ *and* $\max_t C^*_{t,t} \leq c_2$*, for some absolute constants* $c_1, c_2 > 0$*.*

Assumption 2 *The matrix* $\mathbf{\Gamma}^*$ *satisfies:* $\max_{1\leq i\leq d} \gamma_i^* \leq c_3$ *for some absolute constant* $c_3 > 0$*, where* γ_i^* *are the entries of the diagonal matrix* $\mathbf{\Gamma}^*$*.*

Assumptions 1 and 2 are minimal conditions for inference on precision matrices. Furthermore, they imply the conditions needed for clustering consistency derived in Bunea et al. (2018) and discussed in Section 2.2, for n sufficiently large. Their work only requires that $\lambda_{\min}(\mathbf{C}^*)$ is bounded from below by a sequence that converges to zero, as soon as $\|\mathbf{\Sigma}^*\|_{\max}$ and $\|\mathbf{\Gamma}^*\|_{\max}$ are bounded. This is strengthened by our assumptions. In general, a constant lower bound on $\mathbf{C}^*$ is standard in any inference on graphical models and is needed to show the asymptotic normality of the estimator introduced above (Ren et al., 2013; Jankova and van de Geer, 2014; Janková and van de Geer, 2017).

4.2. Asymptotic Normality via Berry-Esseen-type Bounds

4.2.1. Results for the Cluster-Average Graph

In the section, we show that the estimators $\widetilde{\Omega}_{t,k}$ given by (26) are asymptotically normal, for all $t < k$. We define the sparsity of the cluster-average graph as $s_1 \in \mathbb{N}$ such that

$$\max_{1\leq j\leq K} \sum_{k=1}^{K} \mathbb{I}(\Omega^*_{j,k} \neq 0) \leq s_1.$$

Recall that the estimators (22) and (24) depend on the tuning parameters λ and λ'. In the following theorem, we choose $\lambda \asymp \lambda' \asymp \sqrt{\frac{\log(K\vee n)}{n}}$. For notational simplicity, we use C to denote a generic constant, the value of which may change from line to line.

Theorem 3 *If Assumptions 1 and 2 hold, we have*

$$\max_{1\leq t<k\leq K} \sup_{x\in\mathbb{R}} \left|\mathbb{P}(\widehat{T}_{t,k} \leq x) - \Phi(x)\right| \leq \frac{C}{(d\vee n)^3} + \frac{Cs_1 \log(K\vee n)}{n^{1/2}} + \frac{C}{(K\vee n)^3} \tag{33}$$

where $\widehat{T}_{t,k} = \frac{n^{1/2}(\widetilde{\Omega}_{t,k} - \Omega^*_{t,k})}{\widehat{s}_{tk}}$*,* $\widehat{s}^2_{tk} = \widehat{\Omega}^2_{t,k} + \widehat{\Omega}_{t,t}\widehat{\Omega}_{k,k}$ *and* C *is a positive constant.*

Theorem 3, proved in Appendix A, gives the rate of the normal approximation of the distribution of the scaled and centered entries $\widetilde{\Omega}_{t,k}$. The right hand side in (33) is non-asymptotic and is valid for each K, n and d. Its first, small, term is the price to pay for having first used the data for clustering, and it is dominated by the other two terms. From this perspective, the clustering step is the least taxing, as long as we can ensure

its consistency, which in turn can be guaranteed under the minimal assumptions 1 and 2 already needed for the remaining steps.

The second, and dominant, term regards the normal approximation of the distribution of

$$n^{1/2}(\widetilde{\Omega}_{t,k} - \Omega^*_{t,k}). \tag{34}$$

Specifically, as an intermediate step, Proposition 10 in Appendix A shows that the difference between the c.d.f. of (34), scaled by $s_{tk} = \sqrt{\Omega^{*2}_{t,k} + \Omega^*_{t,t}\Omega^*_{k,k}}$, and that of a standard Gaussian random variable is bounded by $\frac{s_1 \log(K \vee n)}{n^{1/2}}$. Therefore, asymptotic normality holds as soon as this quantity converges to zero, which agrees with the weakest sparsity conditions for Gaussian graphical model inference in the literature (Ren et al., 2013; Janková and van de Geer, 2017). In addition, the asymptotic variance s^2_{tk} agrees with the minimum variance bound in Gaussian graphical models (Janková and van de Geer, 2017). Thus, inference based on the empirical risk function (16) does not lead to any asymptotic efficiency loss. Unlike the previous works, we do not require the bounded operator norm condition, $\lambda_{\max}(\boldsymbol{S}^*) \leq C$. This condition is avoided in our analysis by using a more convenient empirical risk function (16), as opposed to the log-likelihood in Jankova and van de Geer (2014), and a CLIME-type initial estimator (24) satisfying (61), as opposed to the node-wise Lasso estimator in Janková and van de Geer (2017).

The last term in the normal approximation is $O((K \vee n)^{-3})$ which is dominated by the second one, and is associated with the replacement of the true variance s^2_{tk} by the estimate $\widehat{s}^2_{tk}$. Finally, we note that the powers of the first and the third term in the right hand side of (33) can be replaced by $2 + \delta$, for any $\delta > 0$, and a change in this power also changes the associated constant C in the term $\frac{C s_1 \log(K \vee n)}{n^{1/2}}$. As shown in Theorem 5, to obtain valid FDR control, we need $K^2/(K \vee n)^{2+\delta} = o(1)$, which holds for any $\delta > 0$. For simplicity, we choose $\delta = 1$ which gives the power 3.

4.2.2. Results for the Latent Variable Graph

In this section we show that the estimators $\widetilde{\Theta}_{t,k}$ given by (32) are asymptotically normal, for all $t < k$. We define the sparsity of the latent graph as $s_0 \in \mathbb{N}$ such that

$$\max_{1 \leq j \leq K} \sum_{k=1}^{K} \mathbb{I}(\Theta^*_{j,k} \neq 0) \leq s_0.$$

Inference for the estimator $\widetilde{\Theta}_{tk}$ follows the general approach outlined in Section 3.1. We prove in Proposition 17 in Appendix B that

$$n^{1/2}(\widetilde{\Theta}_{t,k} - \Theta^*_{t,k}) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} \Theta^*_{t,t} \mathbf{v}_t^{*T} (\bar{\mathbf{C}}^{(i)} \boldsymbol{\Theta}^*_{\cdot k} - \mathbf{e}_k) + o_p(1), \tag{35}$$

where $\mathbf{v}^*_t$ is a K-dimensional vector with $(\mathbf{v}^*_t)_t = 1$ and $(\mathbf{v}^*_t)_{-t} = -\mathbf{w}^*_t$ with $\mathbf{w}^*_t = (\mathbf{C}^*_{-t,-t})^{-1}\mathbf{C}^*_{-t,t}$. and $\bar{\mathbf{C}}^{(i)}$ is defined in (30). The terms of the sum in display (35) are mean zero random variables, and their variance is

$$\sigma^2_{tk} = \mathbb{E}(\Theta^*_{t,t} \mathbf{v}_t^{*T} (\bar{\mathbf{C}}^{(i)} \boldsymbol{\Theta}^*_{\cdot k} - \mathbf{e}_k))^2,$$

which does not have an explicit closed form, unlike the asymptotic variance of the estimates of the entries of $\mathbf{\Omega}^*$. However, we show in Proposition 23 in Appendix B that σ_{tk}^2 admits an approximation that is easy to estimate:

$$\left|\sigma_{tk}^2 - [(\Theta_{t,k}^*)^2 + \Theta_{k,k}^*\Theta_{t,t}^*]\right| \lesssim \frac{s_0}{m},$$

where $m = \min_{1\le k\le K} |G_k^*|$. Guided by this approximation, we estimate σ_{tk}^2 by

$$\widehat{\sigma}_{tk}^2 = \widehat{\Theta}_{t,k}^2 + \widehat{\Theta}_{k,k}\widehat{\Theta}_{t,t}.$$

When all clusters have the equal size, we obtain $K = d/m$. Thus the $O(\frac{s_0}{m})$ terms can be ignored asymptotically in the sense that $\frac{s_0}{m} = \frac{s_0 K}{d} \le \frac{K^2}{d} = o(1)$, when the clusters are approximately balanced, and their number satisfies $K^2 = o(d)$. This is a reasonable assumption in most applied clustering problems. We note that the estimator $\widehat{\sigma}_{tk}^2$ may be inconsistent when the size of some clusters is too small. However, we recall that our ultimate goal is to use these estimators for recovering the sparsity pattern of $\mathbf{\Theta}^*$ under FDR control. To evaluate the sensitivity of our overall procedure to the size of the smallest cluster, we conduct simulation studies in Section 5. The results shows that the proposed method works well as soon as $m > 4$.

The following theorem gives the Berry–Esseen normal approximation bound for the estimators of the entries of the precision matrix corresponding to the latent variable graph.

Theorem 4 *If Assumptions 1 and 2 hold, then*

$$\max_{1\le t<k\le K} \sup_{x\in\mathbb{R}} \left|\mathbb{P}(\widehat{T}_{t,k} \le x) - \Phi(x)\right| \le \frac{C}{(d\vee n)^3} + \frac{C}{(K\vee n)^3} + \frac{Cs_0\log(K\vee n)}{n^{1/2}} + \frac{Cs_0}{m}, \quad (36)$$

where $\widehat{T}_{t,k} = \frac{n^{1/2}(\widetilde{\Theta}_{t,k} - \Theta_{t,k}^*)}{\widehat{\sigma}_{tk}}$ *and* C *is a positive constant.*

Compared to the average graph, the Berry-Esseen bound in (36) contains an additional $\mathcal{O}(\frac{s_0}{m})$ term, stemming from the approximation of the analytically intractable asymptotic variance by an estimable quantity. The proof is deferred to Appendix B.

4.3. Application to Post-clustering FDR Control

Given the edge-wise inferential results for the cluster-average and latent variable graphs established above, we explain in this section how to combine them to control graph-wise inferential uncertainty. Specifically, we view the task of recovering the sparsity pattern as a multiple testing problem by selecting:

$$\mathbf{H}_{0;tk} : \Omega_{t,k}^* = 0 \quad \text{vs.} \quad \mathbf{H}_{1;tk} : \Omega_{t,k}^* \ne 0 \quad \text{for all } 1 \le t < k \le K, \quad (37)$$

for the cluster-average graph, and

$$\mathbf{H}_{0;tk}^{'} : \Theta_{t,k}^* = 0 \quad \text{vs.} \quad \mathbf{H}_{1;tk}^{'} : \Theta_{t,k}^* \ne 0 \quad \text{for all } 1 \le t < k \le K. \quad (38)$$

for the latent variable graph. In the following, we apply the B-Y procedure by Benjamini and Yekutieli (2001) for FDR control in the cluster-average graph. The procedure for latent

variable graph is identical. In this section, we are not claiming to develop a different FDR procedure; rather we make the simple point that if you can obtain asymptotic p-values of d dependent statistics we can combine them in standard ways to control the FDR, which is a direct consequence of the Berry-Esseen bounds derived in Section 4.2.

Define the set of true null hypotheses, $\mathcal{H}_0 := \{(t,k) : 1 \leq t < k \leq K, \text{ such that } \Omega^*_{t,k} = 0\}$, as the set of indices (t,k) for which there is no edge between the nodes t and k. To control the error incurred by multiple testing, we focus on the false discovery rate (FDR), which is the average number of Type I errors relative to the total number of discoveries (Benjamini and Hochberg, 1995). Recall that $\widetilde{\Omega}_{t,k}$ is a consistent and asymptotically normal estimator of $\Omega_{t,k}$. We consider the natural test statistic $\widetilde{W}_{t,k} = n^{1/2}\widetilde{\Omega}_{t,k}/\widehat{s}_{tk}$ for $\mathbf{H}_{0;tk}$, where $\widehat{s}^2_{tk} = \widehat{\Omega}^2_{t,k} + \widehat{\Omega}_{t,t}\widehat{\Omega}_{k,k}$. Given a cutoff $\tau > 0$, the total number of discoveries is

$$R_\tau := \sum_{1 \leq t < k \leq K} \mathbb{I}[|\widetilde{W}_{t,k}| > \tau].$$

Similarly, the number of false positives or false discoveries is given by

$$V_\tau := \sum_{(t,k) \in \mathcal{H}_0} \mathbb{I}[|\widetilde{W}_{t,k}| > \tau].$$

The FDR is formally defined as the expected ratio of V_τ over R_τ,

$$\mathrm{FDR}(\tau) := \mathbb{E}\left[\frac{V_\tau}{R_\tau}\mathbb{I}[R_\tau > 0]\right],$$

where the indicator function is included to remove the trivial case $R_\tau = 0$.

Our goal is to find a data-dependent cutoff τ such that $\mathrm{FDR}(\tau) \leq \alpha + o(1)$ for any given $0 < \alpha < 1$. This is the best one can hope for when, as in our case, the distribution of the test statistics $\widetilde{W}_{t,k}$ is only available asymptotically. The Berry-Esseen type bounds, derived in Theorems 3 and 4, allow us to precisely quantify the price we must pay for the asymptotic approximation and are instrumental to understanding asymptotic FDR control.

In addition, the test statistics $\widetilde{W}_{t,k}$ for different hypotheses are dependent. To allow for the dependence, instead of the standard B-H procedure (Benjamini and Hochberg, 1995), we consider the more flexible B-Y procedure by Benjamini and Yekutieli (2001). The resulting FDR procedure is as follows: reject all hypotheses such that $|\widetilde{W}_{t,k}| \geq \widehat{\tau}$, where

$$\widehat{\tau} := \min\left\{\tau > 0 : \tau \geq \Phi^{-1}\left(1 - \frac{\alpha R_\tau}{2N_{BY}|\mathcal{H}|}\right)\right\} \text{ and } N_{BY} = \sum_{i=1}^{|\mathcal{H}|}\frac{1}{i}, \tag{39}$$

where $|\mathcal{H}| = K(K-1)/2$ is the total number of hypotheses.

Our next result shows when the FDR based on our test statistics is guaranteed to be no greater than α, asymptotically. The proofs can be found in Appendix A.

Theorem 5

1. *Assume that the conditions in Theorem 3 hold. For any* $0 < \alpha < 1$*, we have*

$$\mathrm{FDR}(\widehat{\tau}) \leq \alpha + 2|\mathcal{H}_0|b_n, \tag{40}$$

where $b_n = \frac{C}{(d\vee n)^3} + \frac{C}{(K\vee n)^3} + \frac{Cs_1 \log(K\vee n)}{n^{1/2}}$, *and* $|\mathcal{H}_0|$ *is the number of true null hypotheses in (37).*

2. *Assume that the conditions in Theorem 4 hold. If we define the test statistic as* $\widetilde{W}_{t,k} = n^{1/2}\widetilde{\Theta}_{t,k}/\widehat{\sigma}_{tk}$, *and* $\widehat{\tau}$ *as in (39), we have*

$$\mathrm{FDR}(\widehat{\tau}) \leq \alpha + 2|\mathcal{H}_0'|c_n, \tag{41}$$

where $c_n = \frac{C}{(d\vee n)^3} + \frac{C}{(K\vee n)^3} + \frac{Cs_0 \log(K\vee n)}{n^{1/2}} + \frac{Cs_0}{m}$, *and* $|\mathcal{H}_0'|$ *is the number of true null hypotheses in (38).*

This theorem implies that our method can control the FDR asymptotically, in the sense that $\mathrm{FDR}(\widehat{\tau}) \leq \alpha + o_p(1)$, provided that $s_1|\mathcal{H}_0|\log(K\vee n) = o(n^{1/2})$, for the average graph, and $s_0|\mathcal{H}_0^{'}|\log(K\vee n) = o(n^{1/2})$ for the latent graph. Gaussian graphical model estimation under FDR control was recently studied by Liu (2013). They showed that the B-H procedure can control FDR asymptotically under certain conditions. Their approach is based on the following Cramer-type moderate deviation result using our terminology,

$$\max_{(t,k)\in\mathcal{H}_0} \sup_{0\leq t\leq 2\sqrt{\log K}} \Big| \frac{\mathbb{P}(\widehat{T}_{t,k} \geq t)}{2-2\Phi(t)} - 1 \Big| = o(1), \tag{42}$$

where $\widehat{T}_{t,k}$ is test statistic they proposed for estimation of the Gaussian graphical model structure. The result (42) controls the relative error of the Gaussian approximation within the moderate deviation regime $[0, 2\sqrt{\log K}]$, whereas our result is based on the control of the absolute error via the Berry-Esseen-type Gaussian approximation. One of the main advantages of their result is that the number of clusters is allowed to be $K = o(n^r)$, where r is a constant that can be greater than 1. However, to prove (42), they required that the number of strong signals tends to infinity, that is $|\{(t,k) : \Omega^*_{t,k}/\sqrt{\Omega^*_{k,k}\Omega^*_{t,t}} \geq C\sqrt{\log K/n}\}| \to \infty$, which reduces significantly the parameter space for which inference is valid. In contrast, the aim of this work is the study of pattern recovery without conditions on the signal strength of the entries of the target precision matrices, as in practice it is difficult to assess whether these conditions are met. For completeness, we provide the detailed analysis of the B-H procedure including technical conditions, theoretical results and further discussion of the B-H procedure in Appendix H.

The overall message conveyed by Theorem 5 is that, in the absence of any signal strength assumptions, cluster-based graphical models can still be recovered, under FDR control, provided that the number of clusters K is not very high relative to n, and provided that the clusters are not very small. This further stresses the importance of an initial dimension reduction step in high-dimensional graphical model estimation. For instance, results similar to those of Theorem 3 can be derived along the same lines for the estimation of the sparsity pattern of $\mathbf{\Sigma}^{-1}$, for a generic, unstructured, covariance matrix of $\boldsymbol{X}$, where one replaces K by d throughout, and s_1 is replaced by s, the number of non-zero entries in the $d\times d$ matrix $\mathbf{\Sigma}^{-1}$. Then, the analogue of (41) of Theorem 5 shows that FDR control in generic graphical models, based on *asymptotic approximations of p-values*, cannot generally be guaranteed if $d > n$. Our work shows that extra structural assumptions, for instance those motivated by clustering, do alleviate this problem. The simulation study presented in the next section provides further support to our findings.

5. Numerical Results

This section contains simulations and a real data analysis that illustrate the finite sample performance of the inferential procedures developed in the previous sections for the latent variable graph and cluster-average graph, respectively.

5.1. Synthetic Datasets

In this subsection, we demonstrate the effectiveness of the FDR control procedures on synthetic datasets. We consider two settings $(n, d) = (800, 400)$ and $(500, 1000)$, and in each setting we vary the value of K and m. The error variable $\boldsymbol{E}$ is sampled from the multivariate normal distribution with covariance $\boldsymbol{\Gamma}^*$ whose entries γ_i^* are generated from $U[0.25, 0.5]$. Recall that the latent variable $\boldsymbol{Z}$ follows from $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbf{C}^*)$. We consider three different models to generate the graph structure of $\boldsymbol{Z}$. Once the graph structure is determined, the corresponding adjacency matrix $\mathbf{W}$ is found, and the precision matrix $\boldsymbol{\Theta}^* = \mathbf{C}^{*-1}$ is taken as $\boldsymbol{\Theta}^* = c\mathbf{W} + (|\lambda_{\min}(\mathbf{W})| + 0.2)\mathbf{I}$, where $c = 0.3$ when $d = 400$ and $c = 0.5$ when $d = 1000$. Finally, we assign the cluster labels for all variables so that all clusters have approximately equal size, which gives us the matrix $\mathbf{A}$. Given $\mathbf{A}, \boldsymbol{Z}$ and $\boldsymbol{E}$, we can generate $\boldsymbol{X}$ according to the model (1).

We consider the following three generating models for the graph structure of $\boldsymbol{Z}$:

- *Scale-Free Graph* – The Scale-Free model is a generative model for network data, whose degree distribution follows a power law. To be concrete, we generate the graph one node at a time, starting with a 2 node chain. For nodes $3, \ldots, K$, node t is added and one edge is added between t and one of the $t-1$ previous nodes. Denoting by k_i the current degree of node i in the graph, the probability that node t and node i are connected is $p_i = k_i/(\sum_i k_i)$. The number of edges in the resulting graph is always K.

- *Hub Graph* – The K nodes of the graph are partitioned evenly into groups of size N. Within each group, one node is selected to be the group hub, and an edge is added between it and the remainder of its group. N is either 5 or 6 depending upon the choice of K. The number of edges in the graph is $K(N-1)/N$, so for $K = 100$ with $N = 5$, the number of edges in the resulting graph is 80.

- *Band3 Graph* – This model generates a graph with a Toeplitz adjacency matrix. There is an edge between node i and node j if and only if $|i - j| \leq B$, where we set $B = 3$ in this scenario. In general, the number of edges in a band graph with K nodes is given by $BK - \frac{3}{2}B^2 + \frac{5}{2}B$. So, for $K = 100$ and $B = 3$, the number of edges in the graph is 294.

Recall that $\bar{\boldsymbol{X}} \sim \mathcal{N}(0, \boldsymbol{S}^*)$, where $\boldsymbol{S}^*$ is defined in (15). To determine the structure of the average graph, we numerically compute $\boldsymbol{S}^{*-1}$ and threshold the matrix at 10^{-8}.

We examine the empirical FDR of our procedures on some synthetic datasets. The following protocol is followed in all the experiments:

1. Generate the graph structure of $\boldsymbol{Z}$ as specified above.

2. Simulate n observations from our model (1).

3. Estimate the cluster partition $\widehat{G}$. For computational convenience, we apply the FORCE algorithm (Eisenach and Liu, 2019) with known K.

4. Construct the test statistic $\widetilde{W}_{t,k}$ defined in Section 4.3. The regularization parameters λ and λ' are chosen by 5-fold cross validation.

5. Find the FDR cutoff (39) at level α; we consider three cases $\alpha = 0.05, 0.1, 0.2$.

The simulation is repeated 100 times. To compare with our Benjamini-Yekutieli based FDR procedure, we also report the empirical FDR based on the more classical Benjamini-Hochberg procedure. That is, we apply the same procedures 1-4, but in step 5 we replace the FDR cutoff in (39) with the Benjamini-Hochberg (B-H) cutoff, i.e.,

$$\widehat{\tau}_{BH} := \min\left\{\tau > 0 : \tau \geq \Phi^{-1}\left(1 - \frac{\alpha R_\tau}{2|\mathcal{H}|}\right)\right\}.$$

Table 1 compares the empirical FDR based on our method with the B-H procedure under different m, K settings when $d = 400$. When m is relatively large (e.g., $m = 20$), both methods can control FDR on average, although our method is relatively more conservative. As expected, the FDR control problem becomes more challenging for large K and small m. In this case the graph contains more nodes and each cluster contains fewer variables. We observe that when $m = 5$ our method can still control FDR reasonably well but the B-H method produces empirical FDR far beyond the nominal level, especially for hub graphs. The inferior performance of the B-H procedure is due to the fact that the dependence among the test statistics is not accounted for, demonstrating that the B-Y procedure is indeed necessary at least in the current simulation settings. Finally, we examine the empirical power of the FDR procedure under each scenario, which is defined as

$$\text{Average}\left[\sum_{(t,k)\in\mathcal{H}_1} \frac{\mathbb{I}[\widetilde{W}_{t,k} \geq \widehat{\tau}]}{|\mathcal{H}_1|}\right],$$

where $\mathcal{H}_1$ is the set of alternative hypothesis. Table 2 gives the empirical power of our FDR procedure, and the B-H procedure when $d = 400$. It shows that our procedure and the B-H procedure have very high power in all scenarios. The same phenomenon is observed when d is large, i.e., $d = 1000$; see Tables 3 and 4. In summary, our proposed procedure can identify most of the signals in the graph while keeping FDR well controlled.

Remark 6 (Inexact Recovery) *While our FORCE algorithm guarantees the estimated number of clusters is always K since we use the true value K as the input, the recovered partition $\widehat{G}$ may still differ from the true partition G^*. In order to compare our results to the ground-truth, we need to find an "aligned" version of $\mathbf{\Theta}^*$ (or $\mathbf{\Omega}^*$). Specifically, we calculate the mapping $f : [K] \to [K]$ defined by*

$$f(k) = \underset{l\in[K]}{\text{argmax}}\, |\widehat{G}_k \cap G_l^*|,$$

and then construct the matrix $\mathbf{\Theta}_A^*$, defined entry-wise by

$$(\Theta_A^*)_{s,t} := (\Theta^*)_{f(s),f(t)}.$$

Once we have obtained the "aligned" ground-truth matrix $\mathbf{\Theta}^*_A$, we can proceed with computing the metric of interest (FDR and power). Although the preceding discussion is in terms of $\mathbf{\Theta}^*$, the same procedure applies to $\mathbf{\Omega}^*$.

Remark 7 (Discussion of the B-H procedure) *Based on the results in Tables 3 and 4, it is clear that the power of both our procedure and the B-H procedure are satisfactory. However, for both setups ($d = 400, 1000$), the B-H procedure leads to the inflated FDR relative to the nominal level. While Appendix H shows that the B-H procedure can control FDR asymptotically under certain conditions, it seems in numerical examples the dependence among the test statistics leads to substantial errors in FDR control that are indeed not ignorable.*

				$\boldsymbol{\alpha = 0.05}$			$\boldsymbol{\alpha = 0.1}$			$\boldsymbol{\alpha = 0.2}$		
		K	m	Scalefree	Band3	Hub	Scalefree	Band3	Hub	Scalefree	Band3	Hub
B-Y Based Procedure	Latent	80	5	1.16%	1.42%	5.99%	2.01%	2.64%	7.60%	4.01%	4.73%	10.33%
		66	6	0.93%	1.03%	1.08%	1.99%	1.98%	1.96%	3.73%	3.68%	4.00%
		50	8	1.16%	0.99%	1.09%	2.20%	1.77%	1.98%	3.90%	3.49%	3.73%
		20	20	0.99%	0.88%	1.30%	1.81%	1.66%	2.26%	3.75%	3.38%	4.71%
	Grp. Av.	80	5	1.15%	1.40%	6.29%	2.14%	2.67%	7.89%	4.00%	4.70%	10.43%
		66	6	0.91%	1.04%	1.04%	1.85%	2.00%	2.14%	3.70%	3.71%	3.83%
		50	8	1.16%	1.00%	1.11%	2.22%	1.80%	1.94%	3.98%	3.49%	3.61%
		20	20	0.94%	0.88%	1.23%	1.81%	1.68%	2.26%	3.65%	3.38%	4.71%
B-H Based Procedure	Latent	80	5	8.23%	9.24%	15.60%	15.01%	16.39%	24.00%	28.16%	28.97%	38.58%
		66	6	7.18%	7.38%	7.31%	14.12%	14.01%	13.78%	26.56%	25.84%	27.28%
		50	8	6.94%	6.75%	6.69%	13.23%	12.73%	12.74%	25.20%	23.62%	26.05%
		20	20	5.43%	4.54%	6.38%	11.09%	8.47%	10.71%	21.46%	17.15%	20.75%
	Grp. Av.	80	5	8.42%	9.26%	15.77%	15.25%	16.43%	24.15%	28.03%	29.02%	38.57%
		66	6	7.21%	7.37%	7.45%	13.99%	13.92%	13.81%	26.19%	25.77%	27.17%
		50	8	6.82%	6.78%	6.67%	13.23%	12.67%	12.89%	25.10%	23.61%	25.91%
		20	20	5.38%	4.51%	6.49%	11.01%	8.52%	10.66%	21.58%	17.09%	20.71%

Table 1: Averaged empirical FDR for synthetic data experiments with $d = 400$ and $n = 800$.

Group Average Procedures Do Not Recover Latent Graphs

In this section we demonstrate through simulation studies that procedures specifically tailored to recovering the latent variable graph are necessary. We do this by using both the methodology for the latent variable graph and the group-average graph to recover the latent graph structure. Because the differences between the latent and group-average graph can be small, we use larger sample sizes than in the previous studies.

The experimental procedure is almost exactly the same as before, but now we hold most of the parameters of the generating distribution fixed and examine the effects of the error variance $\mathbf{\Gamma}^*$ and sample size n on the efficacy of our methodologies for recovering the latent variable graph. In all the experiments, we impose a Band3 structure on the latent graph, set $\mathbf{\Gamma}^* = \gamma \mathbf{I}$, and use $(d, K, m) = (400, 20, 20)$.

				$\boldsymbol{\alpha = 0.05}$			$\boldsymbol{\alpha = 0.1}$			$\boldsymbol{\alpha = 0.2}$		
		K	m	Scalefree	Band3	Hub	Scalefree	Band3	Hub	Scalefree	Band3	Hub
B-Y Based Procedure	Latent	80	5	88.22%	99.99%	99.00%	91.10%	100.00%	99.04%	93.38%	100.00%	99.11%
		66	6	93.62%	100%	100%	95.42%	100%	100%	96.86%	100%	100%
		50	8	97.18%	100%	100%	97.90%	100%	100%	98.47%	100%	100%
		20	20	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Grp. Av.	80	5	88.27%	99.99%	98.93%	91.06%	100.00%	98.98%	93.37%	100.00%	99.07%
		66	6	93.62%	100%	100%	95.40%	100%	100%	96.91%	100%	100%
		50	8	97.18%	100%	100%	97.86%	100%	100%	98.47%	100%	100%
		20	20	100%	100%	100%	100%	100%	100%	100%	100%	100%
B-H Based Procedure	Latent	80	5	95.75%	100.00%	99.18%	97.15%	100.00%	99.23%	98.48%	100.00%	99.26%
		66	6	97.85%	100%	100%	98.78%	100%	100%	99.34%	100%	100%
		50	8	99.27%	100%	100%	99.57%	100%	100%	99.80%	100%	100%
		20	20	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Grp. Av.	80	5	95.82%	100.00%	99.14%	97.14%	100.00%	99.18%	98.53%	100.00%	99.23%
		66	6	97.88%	100%	100%	98.78%	100%	100%	99.34%	100%	100%
		50	8	99.24%	100%	100%	99.59%	100%	100%	99.78%	100%	100%
		20	20	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 2: Averaged FDR power for synthetic data experiments with $d = 400$ and $n = 800$.

				$\boldsymbol{\alpha = 0.05}$			$\boldsymbol{\alpha = 0.1}$			$\boldsymbol{\alpha = 0.2}$		
		K	m	Scalefree	Band3	Hub	Scalefree	Band3	Hub	Scalefree	Band3	Hub
B-Y Based Procedure	Latent	80	12	1.40%	1.52%	1.78%	2.78%	2.81%	3.29%	5.71%	5.09%	6.30%
		66	15	1.23%	1.29%	1.31%	2.62%	2.39%	2.69%	5.08%	4.33%	4.84%
		50	20	1.05%	0.94%	1.26%	2.27%	1.85%	2.79%	4.28%	3.65%	4.58%
		20	50	1.04%	1.03%	1.05%	1.81%	1.96%	2.20%	3.46%	3.26%	4.19%
	Grp. Av.	80	12	1.55%	1.57%	1.78%	2.78%	2.80%	3.26%	5.57%	5.13%	6.13%
		66	15	1.30%	1.30%	1.24%	2.64%	2.35%	2.64%	5.07%	4.34%	4.89%
		50	20	1.09%	0.93%	1.26%	2.32%	1.89%	2.75%	4.29%	3.64%	4.51%
		20	50	1.09%	1.01%	1.05%	1.91%	1.96%	2.14%	3.50%	3.28%	4.25%
B-H Based Procedure	Latent	80	12	10.99%	10.04%	12.00%	20.04%	18.18%	21.47%	35.50%	31.46%	37.24%
		66	15	9.28%	8.20%	9.14%	16.88%	14.94%	16.82%	31.53%	27.75%	31.81%
		50	20	7.73%	6.91%	7.96%	14.74%	13.13%	15.33%	27.03%	23.92%	28.72%
		20	50	5.00%	4.85%	6.38%	11.01%	9.38%	11.60%	21.49%	17.00%	23.15%
	Grp. Av.	80	12	10.99%	10.00%	12.00%	20.29%	18.20%	21.38%	35.56%	31.48%	37.10%
		66	15	9.28%	8.24%	8.99%	17.03%	14.87%	16.87%	31.49%	27.82%	31.63%
		50	20	7.60%	6.93%	7.94%	14.61%	13.13%	15.40%	27.12%	23.94%	28.71%
		20	50	4.95%	4.91%	6.38%	10.97%	9.35%	11.60%	21.36%	16.94%	23.41%

Table 3: Averaged empirical FDR for synthetic data experiments with $d = 1000$ and $n = 500$.

First we note that if $\mathbf{\Gamma}^* = \mathbf{0}$, the two graph structures are actually the same – it is only as we introduce error into the observed variables that the latent variable and group-average graphs begin to differ. In Tables 5 and 6 we see that as we increase the error variance, the the group averages methodology applied to the latent graph is unable to control the Type I error rate and the FDR at the desired level. Further, by increasing the sample size n,

				$\boldsymbol{\alpha = 0.05}$			$\boldsymbol{\alpha = 0.1}$			$\boldsymbol{\alpha = 0.2}$		
		K	m	Scalefree	Band3	Hub	Scalefree	Band3	Hub	Scalefree	Band3	Hub
B-Y Based Procedure	Latent	80	12	71.42%	99.95%	100.00%	77.14%	99.97%	100.00%	82.29%	99.98%	100.00%
		66	15	82.98%	99.98%	100.00%	86.32%	99.99%	100.00%	90.02%	100.00%	100.00%
		50	20	90.45%	99.99%	100.00%	93.04%	99.99%	100.00%	94.86%	100.00%	100.00%
		20	50	99.95%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Grp. Av.	80	12	71.54%	99.95%	100.00%	77.13%	99.97%	100.00%	82.25%	99.98%	100.00%
		66	15	82.74%	99.98%	100.00%	86.26%	99.99%	100.00%	90.08%	100.00%	100.00%
		50	20	90.45%	99.99%	100.00%	93.00%	99.99%	100.00%	94.76%	100.00%	100.00%
		20	50	99.95%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
B-H Based Procedure	Latent	80	12	87.15%	99.99%	100.00%	91.33%	100.00%	100.00%	94.66%	100.00%	100.00%
		66	15	93.57%	100.00%	100.00%	95.63%	100.00%	100.00%	97.38%	100.00%	100.00%
		50	20	96.27%	100.00%	100.00%	97.78%	100.00%	100.00%	98.94%	100.00%	100.00%
		20	50	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	Grp. Av.	80	12	87.22%	99.99%	100.00%	91.33%	100.00%	100.00%	94.58%	100.00%	100.00%
		66	15	93.52%	100.00%	100.00%	95.65%	100.00%	100.00%	97.35%	100.00%	100.00%
		50	20	96.33%	100.00%	100.00%	97.78%	100.00%	100.00%	98.94%	100.00%	100.00%
		20	50	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 4: Averaged FDR power for synthetic data experiments with $d = 1000$ and $n = 500$.

we observe that the performance of the group average procedures applied to recovering the latent graph gets worse – this is as expected because with increased n, the tests become accurate, but are by construction estimating the wrong graph!

		L.V. Test Procedure				**G.A. Test Procedure**			
n	γ	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
1600	0.5	0.56%	2.67%	5.83%	10.22%	0.61%	3.00%	6.39%	10.39%
	1.0	0.78%	3.11%	5.22%	10.61%	0.78%	3.83%	6.89%	13.39%
	1.5	0.61%	2.89%	5.11%	8.72%	0.94%	3.33%	6.22%	10.50%
	2.0	1.00%	3.39%	6.39%	11.06%	1.39%	4.78%	7.61%	12.44%
	2.5	0.56%	3.50%	6.17%	11.22%	1.89%	5.33%	9.17%	15.28%
	3.0	1.72%	5.11%	8.61%	14.28%	3.00%	8.56%	11.89%	17.61%
	3.5	1.06%	3.83%	6.78%	11.72%	3.39%	8.44%	12.94%	18.44%
	4.0	1.50%	4.33%	8.33%	12.83%	5.22%	10.94%	14.22%	20.78%
3200	0.5	0.28%	1.89%	3.22%	7.50%	0.44%	2.17%	4.00%	8.11%
	1.0	0.44%	3.00%	5.11%	9.44%	1.11%	4.22%	7.83%	14.17%
	1.5	0.44%	2.72%	4.50%	9.56%	1.78%	5.61%	9.00%	13.61%
	2.0	0.89%	3.56%	6.78%	11.61%	2.94%	7.50%	11.33%	16.78%
	2.5	1.06%	3.61%	6.06%	10.67%	2.78%	9.67%	13.44%	19.00%
	3.0	1.50%	5.44%	8.39%	14.33%	7.72%	13.78%	18.00%	22.78%
	3.5	1.67%	5.44%	9.72%	15.72%	7.61%	13.83%	18.22%	22.89%
	4.0	2.78%	5.83%	10.06%	16.28%	11.28%	17.11%	22.22%	27.06%

Table 5: Averaged Type I for the latent variable graph using both the latent variable and group averages methodology.

		L.V. Test Procedure				**G.A. Test Procedure**			
n	γ	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
1600	0.5	0.83%	3.81%	4.95%	11.44%	0.41%	3.23%	5.51%	12.57%
	1.0	0.62%	3.23%	6.07%	10.11%	1.03%	4.00%	7.51%	14.13%
	1.5	0.21%	3.61%	5.88%	10.61%	1.23%	4.76%	6.25%	12.25%
	2.0	1.23%	3.81%	6.43%	13.51%	1.64%	5.88%	9.26%	14.74%
	2.5	0.83%	4.00%	7.16%	12.25%	3.23%	5.88%	10.95%	16.52%
	3.0	2.24%	6.61%	10.11%	16.81%	3.23%	9.94%	15.19%	20.66%
	3.5	1.23%	4.95%	6.98%	13.36%	3.81%	10.78%	15.79%	22.08%
	4.0	1.64%	4.95%	8.40%	15.19%	6.25%	14.29%	17.95%	24.88%
3200	0.5	0.41%	2.44%	4.00%	7.51%	0.41%	2.64%	4.19%	8.75%
	1.0	0.83%	3.42%	6.80%	10.61%	1.03%	4.57%	8.05%	16.38%
	1.5	0.41%	3.23%	5.70%	9.09%	2.64%	7.34%	10.45%	16.81%
	2.0	0.62%	4.38%	6.98%	12.25%	4.00%	9.77%	13.36%	20.27%
	2.5	1.03%	4.57%	7.51%	11.60%	4.00%	12.09%	16.81%	23.44%
	3.0	1.84%	7.51%	10.45%	16.38%	9.94%	17.81%	22.95%	28.14%
	3.5	1.64%	6.98%	11.28%	18.92%	10.95%	18.23%	23.08%	27.93%
	4.0	4.38%	8.57%	11.60%	18.92%	15.94%	22.95%	27.16%	31.82%

Table 6: Averaged empirical FDR error rates for the latent variable graph using both the latent variable and group averages methodology.

5.2. fMRI Dataset

The study of brain relationships in humans via modern neuroimaging techniques has attracted an enormous amount of scientific interest in recent years. One fundamental goal in these studies is to understand the functional communication between brain regions, which plays a key role in complex cognitive processes. To study the functional connectivity network, Power et al. (2011) partitioned the human brain into regions of interest (ROIs) and represented each ROI by a node in a graph. They identified several "subgraphs" of highly related nodes (i.e, clusters), which can represent the major functional systems of the brain. One of the main goals in their work is to understand the relationship between different subgraphs or clusters. To this end, they estimated the relationship between clusters by thresholding the correlation matrix of all nodes in the same clusters in an adhoc way. The analysis of "network of clusters" can be conducted under a rigorous statistical framework by using the proposed cluster-based graphical models. In the section, we will apply the proposed method to study the dependence structure among functional systems of the brain.

As an illustration, we focus on the publicly available resting-state fMRI data from the Neuro-bureau pre-processed repository (Bellec et al., 2015). Specifically, we use the data from patient 1018959, session 1 in the KKI dataset. This fMRI dataset was pre-processed using the Athena pipeline and mapped to T1 MNI152 coordinate space. We choose this dataset to make our experiments easily reproducible, as the data are available pre-processed using standard alignment and denoising procedures. In a recent work, Luo (2014) estimated the fMRI network in a similar clustering model by an iterative procedure that maximizes the likelihood function with respect to the unknown cluster assignment and the precision

matrix. However, their optimization problem is non-convex due to estimating the unknown cluster assignment and the algorithm is computationally intensive in high dimension.

Following Power et al. (2011), we directly extract the 264 ROIs, which gives us $d = 264$ mean activities across $n = 148$ time periods. Since the FORCE algorithm requires to know the number of clusters in advance, we first apply the CORD algorithm (Bunea et al., 2018) to estimate the number of clusters, which gives us $K = 53$. Then we reapply the FORCE algorithm with $K = 53$ to obtain the corresponding clusters. Recall that the goal is to analyze the dependence structure among functional systems of the brain. Under the latent variable model (1), we treat the measurement of ROIs as the observed variable $\boldsymbol{X}$ and the underlying brain function as the latent variable $\boldsymbol{Z}$. Thus, the relationship between brain functional systems can be modeled by the latent variable graph. Using the FDR control procedures with $\alpha = 0.01$, the estimated latent variable graph is shown in Figure 1. As a comparison, we also show the cluster average graph in Figure 2. However, in this example the biological meaning of the cluster-average variables $\bar{\boldsymbol{X}}$ is unclear so that the interpretation of the cluster average graph seems difficult.

For purposes of clarity, we only display the nodes and connections corresponding to the 10 largest groups in these two figures. The groups are colored according to the functional network the majority of their nodes belong to as given in Power et al. (2011). It is known that the default mode may comprise multiple interacting subsystems (Andrews-Hanna et al., 2010). Thus our graph contains multiple nodes for the default mode, where each node may represent different subsystems. To demonstrate the difference between the latent variable graph and the cluster average graph, we focus on three regions, default mode (pink), dorsal attention (brown) and salience (yellow). In the latent variable graph, dorsal attention and salience are connected and both of them are strongly connected with many subsystems of the default mode. This dependence structure is supported by the neuroscience theory that "the default mode, engaged during rest and internally directed tasks, should exhibit anticorrelation with networks engaged during externally directed tasks, such as the dorsal attention network and the salience network" (Zhou et al., 2017). Moreover, this finding is also consistent with many existing empirical results such as Fox et al. (2005); Fransson (2005); Smith et al. (2009); Raichle (2015). However, in the cluster-averages graph, dorsal attention is conditionally independent of any subsystems of the default mode and salience is only loosely connected with the default mode. In summary, the latent variable graph seems to be more biologically meaningful than the cluster-average graph in order to interpret the dependence structure of the functional systems of the brain.

Acknowledgments

We would like to thank the reviewers and editor for their suggestions to improve this paper. Florentina Bunea was partially supported by NSF-DMS 712709. Yang Ning was partially supported by NSF-DMS 1854637. We are grateful to Xi Luo for help with the interpretation of our data analysis results.

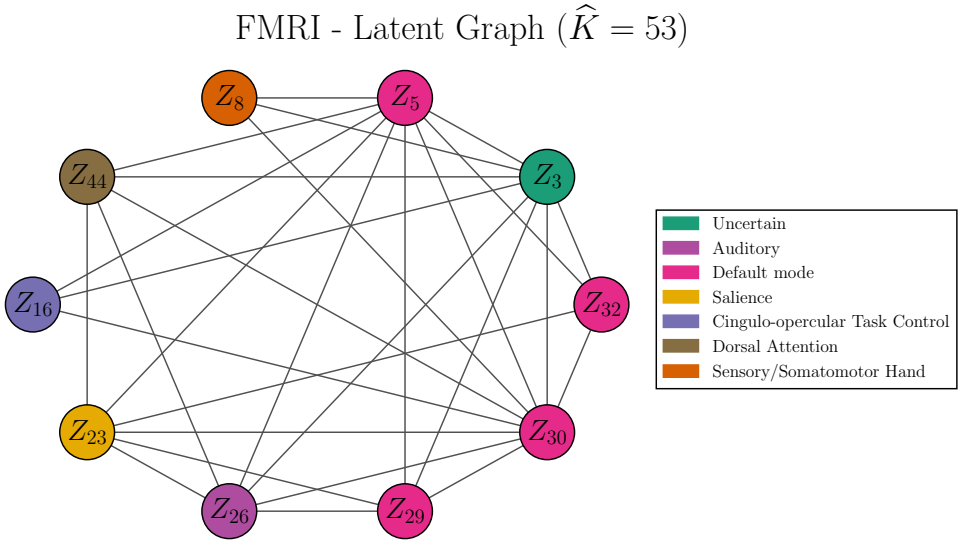

Figure 1: Recovered latent graph structure between 10 largest clusters in fMRI data with FDR level $\alpha = 0.01$ colored according to their functions.

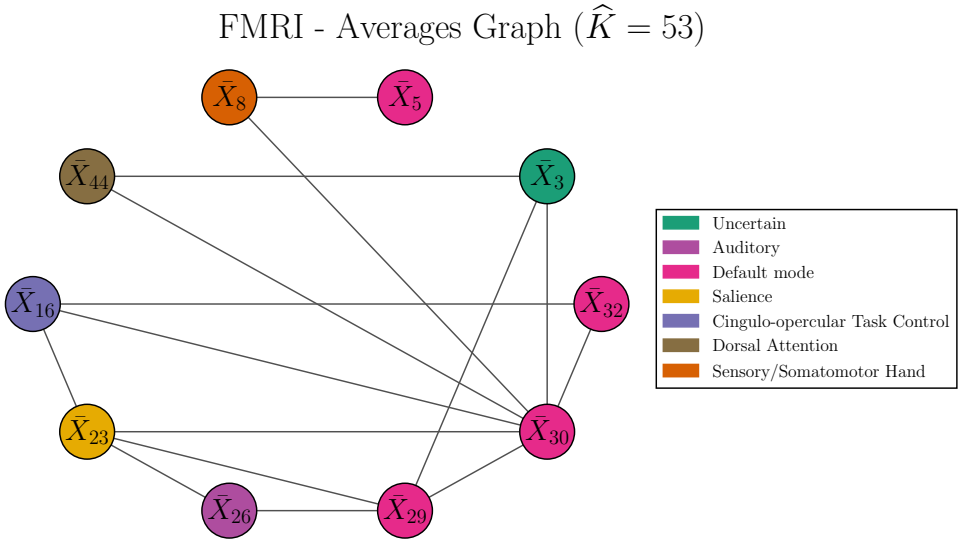

Figure 2: Recovered cluster-averages graph structure between 10 largest clusters in fMRI data with FDR level $\alpha = 0.01$.

Appendix A. Estimation in the Cluster-Average Graph

In this section we provide the proofs of the results needed for establishing the asymptotic normality of the estimators of the entries of $\mathbf{\Omega}^*$. These results make use of the fact that consistent clustering is possible, under our assumptions, as stated below.

Lemma 8 *Let $\mathcal{E} =: \{\widehat{G} = G^*\}$, for $\widehat{G}$ estimated by either the COD or the PECOK algorithm of Bunea et al. (2018). Then, under Assumptions 1 and 2, we have*

$$\mathbb{P}(\mathcal{E}) \geq 1 - \frac{C}{(d \vee n)^3}.$$

The conclusion of this Lemma is proved in Theorem 3, for the COD algorithm, and Theorem 4, for the PECOK algorithm, of Bunea et al. (2018). Lemma 8 allows us to replace $\widehat{G}$ by G^* in all the results below, while incurring a small error, of order $O\left(\frac{C}{(d\vee n)^3}\right)$, which will be shown to be dominated by other error bounds.

Remark 9 *While Assumptions 1 and 2 are made for $\mathbf{C}^*$, they do imply that $c_1 \leq \lambda_{\min}(\boldsymbol{S}^*)$ and $\max_t S^*_{t,t} \leq c_2 + c_3$ holds for $\boldsymbol{S}^*$. Furthermore Lemma 29 implies the same restricted eigenvalue condition on $\lambda_{\min}(\boldsymbol{S}^*)$ as on $\mathbf{C}^*$.*

A.1. Main Proofs for the Cluster-Average Graph Estimators

Before proving Theorem 3 and Claim 1 of Theorem 5, we state two propositions – one regarding the asymptotic normality of $\widetilde{\Omega}_{t,k}$ and the other regarding the convergence rate of the variance estimator. The proofs are deferred until after that of the main result

Proposition 10 (Asymptotic Normality of $\widetilde{\Omega}_{t,k}$) *Under the same conditions as in Theorem 3, we have*

$$\max_{1\leq t<k\leq K} \sup_{x\in\mathbb{R}} \left|\mathbb{P}\Big(\frac{n^{1/2}(\widetilde{\Omega}_{t,k} - \Omega^*_{t,k})}{s_{tk}} \leq x, \mathcal{E}\Big) - \Phi(x)\right| \leq \frac{C}{(K\vee n)^3} + \frac{Cs_1 \log(K\vee n)}{n^{1/2}}.$$

Proposition 11 (Convergence Rate of the Variance Estimator) *Under the same conditions as in Theorem 3, we have*

$$\max_{1\leq t<k\leq K} |\widehat{s}^2_{t,k} - s^2_{t,k}| \leq C\sqrt{\frac{s_1\log(K\vee n)}{n}}, \quad \max_{1\leq t<k\leq K}\left|\frac{\widehat{s}_{t,k}}{s_{t,k}} - 1\right| \leq C\sqrt{\frac{s_1\log(K\vee n)}{n}},$$

with probability at least $1 - (K\vee n)^{-3}$.

Proof of Theorem 3

The proof relies crucially on Proposition 10. Once that result is established, the proof of Theorem 3 follows standard steps as explained below.

Denote $\mathcal{E}' = \{\max_{1\leq t<k\leq K} |\widehat{s}_{tk}/s_{tk} - 1| \leq r\}$, where $r = C\sqrt{\frac{s_1\log(K\vee n)}{n}}$, and let $\bar{\mathcal{E}}'$ signify the complement of the event $\mathcal{E}'$. Let $T_{t,k}$ and $\widehat{T}_{t,k}$ denote the statistics $\frac{n^{1/2}(\widetilde{\Omega}_{t,k} - \Omega^*_{t,k})}{s_{tk}}$ and

$\frac{n^{1/2}(\widetilde{\Omega}_{t,k}-\Omega^*_{t,k})}{\widehat{s}_{tk}}$, respectively. We first consider the bound

$$\begin{aligned}\mathbb{P}\Big(\widehat{T}_{t,k}\le x\Big)-\Phi(x) &\le \mathbb{P}\Big(\widehat{T}_{t,k}\le x,\mathcal{E}',\mathcal{E}\Big)-\Phi(x)+\mathbb{P}(\bar{\mathcal{E}}')+\mathbb{P}(\bar{\mathcal{E}})\\ &= \mathbb{P}\Big(T_{t,k}\le x\frac{\widehat{s}_{tk}}{s_{tk}},\mathcal{E}',\mathcal{E}\Big)-\Phi(x)+\mathbb{P}(\bar{\mathcal{E}}')+\mathbb{P}(\bar{\mathcal{E}}).\end{aligned}$$

Proposition 11 implies $\mathbb{P}(\bar{\mathcal{E}}')\le C(K\vee n)^{-3}$ and Lemma 8 above implies $\mathbb{P}(\bar{\mathcal{E}})\le C(d\vee n)^{-3}$ for some constant C. In addition, for $x\ge 0$,

$$\begin{aligned}&\mathbb{P}\Big(T_{t,k}\le x\frac{\widehat{s}_{tk}}{s_{tk}},\mathcal{E}',\mathcal{E}\Big)-\Phi(x)\le \mathbb{P}\Big(T_{t,k}\le x(1+r),\mathcal{E}\Big)-\Phi(x)\\ &=\Big\{\mathbb{P}\Big(T_{t,k}\le x(1+r),\mathcal{E}\Big)-\Phi(x(1+r))\Big\}+\Big\{\Phi(x(1+r))-\Phi(x)\Big\}.\end{aligned}\tag{43}$$

For the first term, Proposition 10 implies

$$\max_{1\le t<k\le K}\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big(T_{t,k}\le x(1+r),\mathcal{E}\Big)-\Phi(x(1+r))\Big|\lesssim\frac{s_1\log(K\vee n)}{n^{1/2}}.$$

By the mean value theorem, the second term $\Phi(x(1+r))-\Phi(x)=\phi(x(1+tr))xr$, for some $t\in[0,1]$. It is easily seen that $\sup_{x\in\mathbb{R}}\sup_{t\in[0,1]}|\phi(x(1+tr))x|\le C$ for some constant C. Plugging it into (43), we obtain

$$\max_{1\le t<k\le K}\sup_{x\in\mathbb{R}}\Big\{\mathbb{P}\Big(T_{t,k}\le x\frac{\widehat{s}_{tk}}{s_{tk}},\mathcal{E}',\mathcal{E}\Big)-\Phi(x)\Big\}\lesssim\frac{s_1\log(K\vee n)}{n^{1/2}}+r\lesssim\frac{s_1\log(K\vee n)}{n^{1/2}}.\tag{44}$$

When $x<0$, similar to (43), the bound is

$$\begin{aligned}&\mathbb{P}\Big(T_{t,k}\le x\frac{\widehat{s}_{tk}}{s_{tk}},\mathcal{E}',\mathcal{E}\Big)-\Phi(x)\\ &\le\Big\{\mathbb{P}\Big(T_{t,k}\le x(1-r),\mathcal{E}\Big)-\Phi(x(1-r))\Big\}+\Big\{\Phi(x(1-r))-\Phi(x)\Big\}.\end{aligned}$$

Thus (44) holds for $x<0$ as well. Combining these results, we obtain

$$\max_{1\le t<k\le K}\sup_{x\in\mathbb{R}}\Big\{\mathbb{P}\Big(\widehat{T}_{t,k}\le x\Big)-\Phi(x)\Big\}\lesssim\frac{s_1\log(K\vee n)}{n^{1/2}}+\frac{1}{(K\vee n)^3}+\frac{1}{(d\vee n)^3}.$$

Following the similar argument, we can also derive

$$\max_{1\le t<k\le K}\sup_{x\in\mathbb{R}}\Big\{\Phi(x)-\mathbb{P}\Big(\widehat{T}_{t,k}\le x\Big)\Big\}\lesssim\frac{s_1\log(K\vee n)}{n^{1/2}}+\frac{1}{(K\vee n)^3}+\frac{1}{(d\vee n)^3}.$$

This completes the proof.

Proof of Theorem 5, Claim 1

The proof follows verbatim that of Theorem 8.5 in Giraud (2014), with the exception of the fact that exact p-values are replaced by approximate p-values, including the rate

of approximation. We include the full proof for the convenience of the reader. By the definition of the FDR,

$$\mathrm{FDR}(\widehat{\tau}) = \mathbb{E}\Big[\frac{\sum_{(t,k)\in\mathcal{H}_0}\mathbb{I}[|\widetilde{W}_{t,k}|>\widehat{\tau}]\mathbb{I}[R_{\widehat{\tau}}>0]}{R_{\widehat{\tau}}}\Big] = \sum_{(t,k)\in\mathcal{H}_0}\mathbb{E}\Big[\frac{\mathbb{I}[|\widetilde{W}_{t,k}|>\widehat{\tau}]\mathbb{I}[R_{\widehat{\tau}}>0]}{R_{\widehat{\tau}}}\Big]. \quad (45)$$

To handle the $R_{\widehat{\tau}}$ in the denominator, we use the identity $1 = \sum_{i=R_{\widehat{\tau}}}^{\infty}\frac{R_{\widehat{\tau}}}{i(i+1)}$. This implies

$$1/R_{\widehat{\tau}} = \sum_{i=R_{\widehat{\tau}}}^{\infty}\frac{1}{i(i+1)} = \sum_{i=1}^{\infty}\frac{\mathbb{I}[i\geq R_{\widehat{\tau}}]}{i(i+1)}$$

Plugging this into (45) and bringing the expectation inside the summation gives that

$$\begin{aligned}
\mathrm{FDR}(\widehat{\tau}) &= \sum_{(t,k)\in\mathcal{H}_0}\sum_{i=1}^{\infty}\frac{1}{i(i+1)}\mathbb{E}\Big[\mathbb{I}[|\widetilde{W}_{t,k}|>\widehat{\tau}]\mathbb{I}[R_{\widehat{\tau}}>0]\mathbb{I}[i\geq R_{\widehat{\tau}}]\Big] \\
&\leq \sum_{(t,k)\in\mathcal{H}_0}\sum_{i=1}^{\infty}\frac{1}{i(i+1)}\mathbb{E}\Big[\mathbb{I}[|\widetilde{W}_{t,k}|>\Phi^{-1}\Big(1-\frac{\alpha R_{\widehat{\tau}}}{2N_{BY}|\mathcal{H}|}\Big)]\mathbb{I}[R_{\widehat{\tau}}>0]\mathbb{I}[i\geq R_{\widehat{\tau}}]\Big] \\
&\leq \sum_{(t,k)\in\mathcal{H}_0}\sum_{i=1}^{\infty}\frac{1}{i(i+1)}\mathbb{E}\Big[\mathbb{I}[|\widetilde{W}_{t,k}|>\Phi^{-1}\Big(1-\frac{\alpha(i\wedge|\mathcal{H}|)}{2N_{BY}|\mathcal{H}|}\Big)]\Big], \qquad (46)
\end{aligned}$$

where the second line follows from the definition of the FDR cutoff and the last inequality holds since $R_{\widehat{\tau}}\leq(i\wedge|\mathcal{H}|)$. The Berry-Esseen bound in Theorem 3 implies that

$$\mathbb{P}(|\widetilde{W}_{t,k}|>\Phi^{-1}\Big(1-\frac{\alpha(i\wedge|\mathcal{H}|)}{2N_{BY}|\mathcal{H}|}\Big)) \leq \frac{\alpha(i\wedge|\mathcal{H}|)}{N_{BY}|\mathcal{H}|}+2b_n.$$

Thus, it follows that

$$\begin{aligned}
\mathrm{FDR}(\widehat{\tau}) &\leq \sum_{(t,k)\in\mathcal{H}_0}\sum_{i=1}^{\infty}\frac{1}{i(i+1)}\Big(\frac{\alpha(i\wedge|\mathcal{H}|)}{N_{BY}|\mathcal{H}|}+2b_n\Big) \\
&= \alpha\frac{|\mathcal{H}_0|}{|\mathcal{H}|}\Big(\sum_{i=1}^{|\mathcal{H}|}\frac{i}{i(i+1)N_{BY}}+\sum_{i=|\mathcal{H}|+1}^{\infty}\frac{|\mathcal{H}|}{i(i+1)N_{BY}}\Big)+2|\mathcal{H}_0|b_n \\
&= \alpha\frac{|\mathcal{H}_0|}{|\mathcal{H}|}\Big(\sum_{i=1}^{|\mathcal{H}|}\frac{1}{i+1}\frac{1}{N_{BY}}+\frac{|\mathcal{H}|}{|\mathcal{H}|+1}\frac{1}{N_{BY}}\Big)+2|\mathcal{H}_0|b_n \\
&\leq \alpha+2|\mathcal{H}_0|b_n,
\end{aligned}$$

where the last step follows from $\frac{|\mathcal{H}_0|}{|\mathcal{H}|}\leq 1$ and the definition of N_{BY}. This completes the proof of the Theorem 5, Claim 1.

Proof of Proposition 10

Proof The proof is done in two steps. In Step 1, we show that intersected with the event $\mathcal{E}$,

$$n^{1/2}|(\widetilde{\Omega}_{t,k} - \Omega^*_{t,k})/s_{t,k} + \Omega^*_{t,t} h(\mathbf{\Omega}^*_{\cdot k})/s_{t,k}| \leq \frac{s_1 \log(K \vee n)}{n^{1/2}}, \tag{47}$$

with probability at least $1 - C/(K \vee n)^3$ and then use Lemma 16 to obtain the result. To prove (47), we decompose it as

$$\begin{aligned}
&n^{1/2}|(\widetilde{\Omega}_{t,k} - \Omega^*_{t,k}) + \Omega^*_{t,t} h(\mathbf{\Omega}^*_{\cdot k})| \\
&= n^{1/2}|(\widehat{\Omega}_{t,k} - \Omega^*_{t,k}) - \widehat{\Omega}_{t,t}\widehat{h}(\widehat{\mathbf{\Omega}}_{\cdot k}) + \Omega^*_{t,t} h(\mathbf{\Omega}^*_{\cdot k})| \\
&\leq \underbrace{n^{1/2}|(\widehat{\Omega}_{t,k} - \Omega^*_{t,k}) - \Omega^*_{t,t}(h(\widehat{\mathbf{\Omega}}_{\cdot k}) - h(\mathbf{\Omega}^*_{\cdot k}))|}_{\text{I.1}} \\
&\quad + \underbrace{n^{1/2}|\Omega^*_{t,t}(\widehat{h}(\widehat{\mathbf{\Omega}}_{\cdot k}) - h(\widehat{\mathbf{\Omega}}_{\cdot k}))|}_{\text{I.2}} + \underbrace{n^{1/2}|(\widehat{\Omega}_{t,t} - \Omega^*_{t,t})\widehat{h}(\widehat{\mathbf{\Omega}}_{\cdot k})|}_{\text{I.3}}.
\end{aligned}$$

In the following, we study these three terms separately. Recall that $h(\mathbf{\Omega}_{\cdot k}) = \mathbf{v}_t^{*T}(\widehat{\boldsymbol{S}}\mathbf{\Omega}_{\cdot k} - \mathbf{e}_k)$, and $\widehat{h}(\mathbf{\Omega}_{\cdot k}) = \widehat{\mathbf{v}}_t^T(\widehat{\boldsymbol{S}}\mathbf{\Omega}_{\cdot k} - \mathbf{e}_k)$, where $\mathbf{v}_t^*$ is a K-dimensional vector with $(\mathbf{v}_t^*)_t = 1$ and $(\mathbf{v}_t^*)_{-t} = -\mathbf{w}_t^*$ with $\mathbf{w}_t^* = (\boldsymbol{S}^*_{-t,-t})^{-1}\boldsymbol{S}^*_{-t,t}$. Term I.1 reduces to

$$\begin{aligned}
|I.1| &= n^{1/2}|(\widehat{\Omega}_{t,k} - \Omega^*_{t,k}) - \Omega^*_{t,t}\mathbf{v}_t^{*T}\widehat{\boldsymbol{S}}(\widehat{\mathbf{\Omega}}_{\cdot k} - \mathbf{\Omega}^*_{\cdot k})| \\
&\leq n^{1/2}|(\widehat{\Omega}_{t,k} - \Omega^*_{t,k})(1 - \Omega^*_{t,t}(\widehat{\boldsymbol{S}}_{t,t} - \mathbf{w}_t^{*T}\widehat{\boldsymbol{S}}_{-t,t}))| \qquad (48) \\
&\quad + n^{1/2}\Omega^*_{t,t}|(\widehat{\boldsymbol{S}}_{t,-t} - \mathbf{w}_t^{*T}\widehat{\boldsymbol{S}}_{-t,-t})(\widehat{\mathbf{\Omega}}_{-t,k} - \mathbf{\Omega}^*_{-t,k})|. \qquad (49)
\end{aligned}$$

Note that $1/\Omega^*_{t,t} = S^*_{t,t} - \mathbf{w}_t^{*T}\boldsymbol{S}^*_{-t,t}$. The term in (48) can be bounded by

$$\begin{aligned}
&n^{1/2}|(\widehat{\Omega}_{t,k} - \Omega^*_{t,k})\Omega^*_{t,t}(\widehat{S}_{t,t} - S^*_{t,t})| + n^{1/2}|(\widehat{\Omega}_{t,k} - \Omega^*_{t,k})\Omega^*_{t,t}\mathbf{w}_t^{*T}(\widehat{\boldsymbol{S}}_{-t,t} - \boldsymbol{S}^*_{-t,t})| \\
&\leq n^{1/2}|\Omega^*_{t,t}|\|\widehat{\mathbf{\Omega}}_{\cdot k} - \mathbf{\Omega}^*_{\cdot k}\|_1 \max(\|\widehat{\boldsymbol{S}} - \boldsymbol{S}^*\|_{\max}, \|\mathbf{w}_t^{*T}(\widehat{\boldsymbol{S}}_{-t\cdot} - \boldsymbol{S}^*_{-t\cdot})\|_\infty) \\
&\leq \frac{Cs_1 \log(K \vee n)}{n^{1/2}},
\end{aligned}$$

with probability at least $1 - (K \vee n)^{-3}$, by $\lambda_{\max}(\mathbf{\Omega}^*) \leq C$ and the concentration and error bound results in Lemmas 13, 14, 15. The term in (49) can be bounded by

$$n^{1/2}\Omega^*_{t,t}\|\widehat{\boldsymbol{S}}_{t,-t} - \mathbf{w}_t^{*T}\widehat{\boldsymbol{S}}_{-t,-t}\|_\infty\|\widehat{\mathbf{\Omega}}_{-t,k} - \mathbf{\Omega}^*_{-t,k}\|_1 \leq \frac{Cs_1 \log(K \vee n)}{n^{1/2}},$$

with probability at least $1 - (K \vee n)^{-3}$ again by Lemmas 14, 15. Thus, $|\text{I.1}| \leq \frac{s_1 \log(K \vee n)}{n^{1/2}}$ with probability at least $1 - (K \vee n)^{-3}$.

For term I.2, we have

$$\begin{aligned}
|I.2| &= n^{1/2}\Omega^*_{t,t}|(\widehat{\mathbf{v}}_t - \mathbf{v}^*)^T(\widehat{\boldsymbol{S}}\widehat{\mathbf{\Omega}}_{\cdot k} - \mathbf{e}_k)| \\
&\leq n^{1/2}\Omega^*_{t,t}\|\widehat{\mathbf{v}}_t - \mathbf{v}^*\|_1\|\widehat{\boldsymbol{S}}\widehat{\mathbf{\Omega}}_{\cdot k} - \mathbf{e}_k\|_\infty \leq \frac{Cs_1 \log(K \vee n)}{n^{1/2}},
\end{aligned}$$

with probability at least $1-(K\vee n)^{-3}$ by Lemma 15 and the constraint of the CLIME-type estimator.

To control term I.3, first we observe that

$$\begin{aligned}|\widehat{h}(\widehat{\mathbf{\Omega}}_{\cdot k})| &= |\widehat{\mathbf{v}}_t^T(\widehat{\boldsymbol{S}}\widehat{\mathbf{\Omega}}_{\cdot k}-\mathbf{e}_k)|\\ &\le |\mathbf{v}_t^{*T}(\widehat{\boldsymbol{S}}\mathbf{\Omega}^*_{\cdot k}-\mathbf{e}_k)| + |\mathbf{v}_t^{*T}\widehat{\boldsymbol{S}}(\widehat{\mathbf{\Omega}}_{\cdot k}-\mathbf{\Omega}^*_{\cdot k})| + |(\widehat{\mathbf{v}}_t-\mathbf{v}_t^*)^T(\widehat{\boldsymbol{S}}\widehat{\mathbf{\Omega}}_{\cdot k}-\mathbf{e}_k)|\\ &\le \|\mathbf{v}_t^*\|_1\|\widehat{\boldsymbol{S}}\mathbf{\Omega}^*_{\cdot k}-\mathbf{e}_k\|_\infty + \|\widehat{\boldsymbol{S}}_{t,-t}-\mathbf{w}_t^{*T}\widehat{\boldsymbol{S}}_{-t,-t}\|_\infty\|\widehat{\mathbf{\Omega}}_{\cdot k}-\mathbf{\Omega}^*_{\cdot k}\|_1\\ &\quad + \|\widehat{\mathbf{v}}_t-\mathbf{v}_t^*\|_1\|\widehat{\boldsymbol{S}}\widehat{\mathbf{\Omega}}_{\cdot k}-\mathbf{e}_k\|_\infty.\end{aligned}$$

As shown in the proof of Lemma 16, $\|\mathbf{v}_t^*\|_1 \le s_1^{1/2}\|\mathbf{v}_t^*\|_2 \le Cs_1^{1/2}$. The rest of the bounds on the above terms follows easily from Lemmas 14, 15. Thus, we have $|\widehat{h}(\widehat{\mathbf{\Omega}}_{\cdot k})| \le C(s_1\log(K\vee n)/n)^{1/2}$ with high probability. Since

$$|\widehat{\Omega}_{t,t}-\Omega^*_{t,t}| \le C(s_1\log(K\vee n)/n)^{1/2},$$

by Lemma 15, we obtain that $|\text{I.3}| \le \frac{s_1\log(K\vee n)}{n^{1/2}}$ with probability at least $1-(K\vee n)^{-3}$. It is easily seen that $\Omega^*_{t,t} \ge \frac{1}{S^*_{t,t}} \ge C > 0$, see Remark 9. This implies that $s^2_{t,k} = \Omega^*_{t,t}\Omega^*_{k,k} + \Omega^{*2}_{t,k}$ is lower bounded by a positive constant. The proof of (47) is complete.

In step 2, we need to verify that

$$\max_{1\le t<k\le K}\sup_{x\in\mathbb{R}}\Big|\mathbb{P}\Big(\frac{n^{1/2}\Omega^*_{t,t}h(\mathbf{\Omega}^*_{\cdot k})}{s_{tk}} \le x\Big) - \Phi(x)\Big| \le \frac{C}{(K\vee n)^3} + \frac{1}{n^{1/2}},$$

which has been done in Lemma 16. Thus, combining with result (47), we can use the same simple union bound as in the proof of Theorem 3 to obtain the desired result. ■

Proof of Proposition 11

Proof By Lemma 15, under the event $\widehat{G} = G^*$, we have

$$\max_{1\le t,k\le K}|\widehat{\Omega}_{t,k}-\Omega^*_{t,k}| \le \max_{1\le k\le K}\|\widehat{\mathbf{\Omega}}_{\cdot k}-\mathbf{\Omega}^*_{\cdot k}\|_2 \le C_1\sqrt{\frac{s_1\log(K\vee n)}{n}},$$

with probability at least $1-\frac{C_4}{(K\vee n)^3}$. Under this event,

$$\begin{aligned}\max_{1\le t<k\le K}|\widehat{s}^2_{t,k}-s^2_{t,k}| &= \max_{1\le t<k\le K}|\widehat{\Omega}^2_{t,k}+\widehat{\Omega}_{t,t}\widehat{\Omega}_{k,k}-(\Omega^{*2}_{t,k}+\Omega^*_{t,t}\Omega^*_{k,k})|\\ &\le \max_{1\le t<k\le K}|(\widehat{\Omega}_{t,k}-\Omega^*_{t,k})(\widehat{\Omega}_{t,k}+\Omega^*_{t,k})|\\ &\quad + \widehat{\Omega}_{t,t}|\widehat{\Omega}_{k,k}-\Omega^*_{k,k}| + \Omega^*_{k,k}|\widehat{\Omega}_{t,t}-\Omega^*_{t,t}|\\ &\le C_1\sqrt{\frac{s_1\log(K\vee n)}{n}}(4\|\mathbf{\Omega}^*\|_{\max}+\delta)\\ &\le C\sqrt{\frac{s_1\log(K\vee n)}{n}},\end{aligned}$$

for some constant $\delta > 0$ since $\|\mathbf{\Omega}^*\|_{\max} \leq \lambda_{\max}(\mathbf{\Omega}^*) \leq C$. It is easily verified that $\Omega^*_{t,t} \geq \frac{1}{S^*_{t,t}} \geq C > 0$ (see Remark 9). This implies that $s^2_{t,k} = \Omega^*_{t,t}\Omega^*_{k,k} + \Omega^{*2}_{t,k}$ is lower bounded by a positive constant. Thus,

$$\max_{1\leq t<k\leq K} \Big|\frac{\widehat{s}_{t,k}}{s_{t,k}} - 1\Big| \leq \max_{1\leq t<k\leq K} \Big|\frac{\widehat{s}^2_{t,k} - s^2_{t,k}}{s_{t,k}(s_{t,k} + \widehat{s}_{t,k})}\Big| \leq C\sqrt{\frac{s_1 \log(K \vee n)}{n}}.$$

■

A.2. Key Lemmas for Estimators of the Cluster-Average Graph

Remark 12 *In the following proofs, we always assume the event $\mathcal{E} = \{\widehat{G} = G^*\}$ holds. Using a similar argument to that used in the proof of Theorem 3, the following bounds will hold with probability at least $1 - \frac{C}{(K\vee n)^3}$.*

Lemma 13 (Consistency of $\widehat{\boldsymbol{S}}$) *If Assumptions 1 and 2 hold, then with probability greater than $1 - \frac{C}{(K\vee n)^3}$,*

$$\|\widehat{\boldsymbol{S}} - \boldsymbol{S}^*\|_{\max} \leq C\sqrt{\frac{\log(K \vee n)}{n}}$$

for some constant C dependent only on c_1, c_2, and c_3 from Assumptions 1 and 2.

Proof The proof follows from the proof of Theorem 1 in Cai et al. (2011). ■

Lemma 14 (Concentration of Gradient and Hessian) *If Assumptions 1 and 2 hold, then with probability greater than $1 - \frac{C}{(K\vee n)^3}$, we have that*

(a) $\max_{1\leq k\leq K} \|\widehat{\boldsymbol{S}}\mathbf{\Omega}^*_{\cdot k} - \mathbf{e}_k\|_\infty \leq C_1\sqrt{\frac{\log(K\vee n)}{n}}$,

(b) $\max_{1\leq t\leq K} \|\widehat{\boldsymbol{S}}_{t,-t} - \mathbf{w}^{*T}_t \widehat{\boldsymbol{S}}_{-t,-t}\|_\infty \leq C_2\sqrt{\frac{\log(K\vee n)}{n}}$, *and*

(c) $\max_{1\leq t\leq K} \|\mathbf{w}^{*T}_t(\widehat{\boldsymbol{S}}_{-t,-t} - \boldsymbol{S}^*_{-t,-t})\|_\infty \leq C_2\sqrt{\frac{\log(K\vee n)}{n}}$,

for some constants C_1 and C_2 dependent only on c_1, c_2, and c_3 from Assumptions 1 and 2.

Proof We start from the decomposition

$$\widehat{\boldsymbol{S}} - \boldsymbol{S}^* = (\mathbf{A}^{*T}\mathbf{A}^*)^{-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)\mathbf{A}^*(\mathbf{A}^{*T}\mathbf{A}^*)^{-1}.$$

Denoting by $\mathbf{B}^* = \mathbf{A}^{*T}\mathbf{A}^*$, we can write

$$\begin{aligned}
&(\mathbf{A}^{*T}\mathbf{A}^*)^{-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)\mathbf{A}^*(\mathbf{A}^{*T}\mathbf{A}^*)^{-1} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}\Big\{(\mathbf{A}^*\boldsymbol{Z}_i + \boldsymbol{E}_i)(\mathbf{A}^*\boldsymbol{Z}_i + \boldsymbol{E}_i)^T - \mathbf{A}^*\mathbf{C}^*\mathbf{A}^{*T} - \mathbf{\Gamma}^*\Big\}\mathbf{A}^*\mathbf{B}^{*-1} \\
&= \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T - \mathbf{C}^* + \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{E}_i^T\mathbf{A}^*\mathbf{B}^{*-1} + \frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}\boldsymbol{E}_i\boldsymbol{Z}_i^T \\
&\quad + \frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}(\boldsymbol{E}_i\boldsymbol{E}_i^T - \mathbf{\Gamma}^*)\mathbf{A}^*\mathbf{B}^{*-1}. \qquad (50)
\end{aligned}$$

First, note that $\widehat{\boldsymbol{S}}\mathbf{\Omega}^*_{\cdot k} - \mathbf{e}_k = (\widehat{\boldsymbol{S}} - \boldsymbol{S}^*)\mathbf{\Omega}^*_{\cdot k}$, and $||\mathbf{\Omega}^*_{\cdot k}||_2 \leq \lambda_{\max}(\mathbf{\Omega}^*) \leq \lambda_{\min}(\boldsymbol{S}^*)^{-1}$. As seen in Remark 9, we can bound the smallest eigenvalue of $\boldsymbol{S}^*$ from below by $\lambda_{\min}(\mathbf{C}^*)$. We therefore obtain that $||\mathbf{\Omega}^*_{\cdot k}||_2 \leq c_1 + c_3$. Using this bound, we apply Lemma 24, part (d) to control the term

$$\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{Z}_i\boldsymbol{Z}_i^T - \mathbf{C}^*\right)\mathbf{\Omega}^*_{\cdot k}.$$

The remaining terms in (50) can similarly be controlled using Lemma 25 and Lemma 26. Applying the triangle inequality therefore gives

$$\max_{1\leq k\leq K} \|\widehat{\boldsymbol{S}}\mathbf{\Omega}^*_{\cdot k} - \mathbf{e}_k\|_\infty \leq C_1\sqrt{\frac{\log(K\vee n)}{n}}$$

with probability at least $1 - \frac{C}{(K\vee n)^3}$, concluding the proof of claim (a).

For the remaining two claims, we can rewrite

$$\mathbf{w}_t^* = \left(S^*_{t,t} - \boldsymbol{S}^{*T}_{-t,t}(\boldsymbol{S}^*_{-t,-t})^{-1}\boldsymbol{S}^*_{-t,t}\right)\mathbf{\Omega}^*_{-t,t} = \frac{1}{\Omega^*_{t,t}}\mathbf{\Omega}^*_{-t,t}$$

by the block matrix inverse formula. Using Lemma 30, it follows that $||\mathbf{w}_t^*||_2 \leq \lambda_{\min}(\mathbf{\Omega}^*)\max_t S^*_{t,t}$. Then we can see that

$$\begin{aligned}
\max_{1\leq t\leq K} ||\widehat{\boldsymbol{S}}_{t,-t} - \mathbf{w}_t^{*T}\widehat{\boldsymbol{S}}_{-t,-t}||_\infty &= \max_{1\leq t\leq K} ||(\widehat{\boldsymbol{S}}_{t,-t} - \boldsymbol{S}^*_{t,-t}) - \mathbf{w}_t^{*T}(\widehat{\boldsymbol{S}}_{-t,-t} - \boldsymbol{S}^*_{-t,-t})||_\infty \\
&\leq \underbrace{\max_{1\leq t\leq K} ||(\widehat{\boldsymbol{S}}_{t,-t} - \boldsymbol{S}^*_{t,-t})||_\infty}_{\text{(i)}} + \underbrace{\max_{1\leq t\leq K} ||\mathbf{w}_t^{*T}(\widehat{\boldsymbol{S}}_{-t,-t} - \boldsymbol{S}^*_{-t,-t})||_\infty}_{\text{(ii)}}.
\end{aligned}$$

By using Lemma 13, (i) is bounded with high probability. Following the same step as in the proof of (a), we have that $(ii) \leq C_2\sqrt{\frac{\log(K\vee n)}{n}}$, with probability at least $1 - \frac{C}{(K\vee n)^3}$. Part (c) is the same as the term (ii), concluding the proof. ■

Lemma 15 (Consistency of Initial Estimator) *If Assumptions 1 and 2 hold, then,*

(a) $\max_{1\leq k\leq K} \|\widehat{\mathbf{\Omega}}_{\cdot k} - \mathbf{\Omega}^*_{\cdot k}\|_1 \leq C_1 s_1\sqrt{\frac{\log(K\vee n)}{n}}$,

(b) $\max_{1\leq k\leq K}\|\widehat{\boldsymbol{\Omega}}_{\cdot k}-\boldsymbol{\Omega}^*_{\cdot k}\|_2\leq C_1\sqrt{\frac{s_1\log(K\vee n)}{n}}$,

(c) $\max_{1\leq t\leq K}\|\widehat{\mathbf{v}}_t-\mathbf{v}^*_t\|_1\leq C_2 s_1\sqrt{\frac{\log(K\vee n)}{n}}$, *and*

(d) $\max_{1\leq k\leq t\leq K}|(\widehat{\mathbf{v}}_t-\mathbf{v}^*_t)^T\widehat{\boldsymbol{S}}(\widehat{\boldsymbol{\Omega}}_{\cdot k}-\boldsymbol{\Omega}^*_{\cdot k})|\leq C_3\frac{s_1\log(K\vee n)}{n}$,

with probability at least $1-\frac{C_4}{(K\vee n)^3}$. C_1, C_2, C_3, C_4 *and* C_5 *are constants, dependent only upon* c_0, c_1, c_2, *and* c_3 *from Assumptions 1 and 2.*

Proof The proof follows from the same argument as in the proof of Lemma 21 with Lemma 20 replaced by Lemma 14. ■

Lemma 16 (CLT for the Pseudo-Score Function) *Recall that* $s^2_{tk}=\Omega^{*2}_{t,k}+\Omega^*_{t,t}\Omega^*_{k,k}$. *Let* F_n *denote the CDF of* $n^{1/2}\mathbf{v}^{*T}_t(\widehat{\boldsymbol{S}}\boldsymbol{\Omega}^*_{\cdot k}-\mathbf{e}_k)/(s_{tk}/\Omega^*_{t,t})$. *If Assumptions 1 and 2 hold, then*

$$\max_{1\leq t<k\leq K}\sup_{x\in\mathbb{R}}|F_n(x)-\Phi(x)|\leq C(n^{-1/2}+(d\vee n)^{-3}),$$

where C *is a constant dependent only upon* c_0, c_1, *and* c_2.

Proof The proof is very similar to that of Lemma 22, so we only summarize the key steps here. For the cluster averages score function, we have the similar bound that

$$F_n(x)-\Phi(x)\leq\widetilde{F}_n(x)-\Phi(x)+\mathbb{P}(\bar{\mathcal{E}}),$$

where $\mathcal{E}$ is the event $\widehat{G}=G^*$, and $\widetilde{F}_n(x)$ is the CDF of $n^{-1/2}\sum_{i=1}^n\mathbf{v}^{*T}_t(\bar{\boldsymbol{X}}_i\bar{\boldsymbol{X}}_i^T\boldsymbol{\Theta}^*_{\cdot k}-\mathbf{e}_k)/(s_{tk}/\Theta^*_{tt})$. Lemma 8 implies that $\mathbb{P}(\bar{\mathcal{E}})\leq(d\vee n)^{-3}$.

As in Lemma 22, we can verify the Lyapunov condition to control $\widetilde{F}_n(x)-\Phi(x)$. Lastly, from the Isserlis' theorem, we get that

$$\mathrm{Var}(\boldsymbol{v}_1^T\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\boldsymbol{v}_2)=(\boldsymbol{v}_1^T\boldsymbol{S}^*\boldsymbol{v}_1)(\boldsymbol{v}_2^T\boldsymbol{S}^*\boldsymbol{v}_2)+(\boldsymbol{v}_1^T\boldsymbol{S}^*\boldsymbol{v}_2)^2,$$

for any vector $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Using this, it is straightforward to shown

$$\mathbb{E}[\mathbf{v}^{*T}_t(\bar{\boldsymbol{X}}_i\bar{\boldsymbol{X}}_i^T\boldsymbol{\Theta}^*_{\cdot k}-\mathbf{e}_k)^2]=(s_{tk}/\Theta^*_{tt})^2,$$

concluding the proof. ■

Appendix B. Estimation in the Latent Variable Graph

This section contains the proofs which establish the asymptotic normality of the estimator of $\boldsymbol{\Theta}^*$.

B.1. Main Proofs for the Latent Variable Graph Estimators

As in Section B we develop the proofs of Theorem 4 and Claim 2 of Theorem 5 by first establishing two supporting propositions.

Proposition 17 (Asymptotic Normality of $\widetilde{\Theta}_{t,k}$) *Under the same conditions as in Theorem 4, we get*

$$\max_{1\le t<k\le K}\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{n^{1/2}(\widetilde{\Theta}_{t,k}-\Theta^*_{t,k})}{\sigma_{t,k}}<x,\mathcal{E}\right)-\Phi(x)\right|\le\frac{C}{(K\vee n)^3}+\frac{Cs_0\log(K\vee n)}{n^{1/2}}.$$

Proof The proof follows all the steps of Proposition 10, with Lemma 13 replaced by Lemma 19, 14 by 20, 15 by 21 and 16 by 22. ■

Proposition 18 (Convergence Rate of the Variance Estimator) *Under the same conditions as in Theorem 4, we get*

$$\max_{1\le t<k\le K}|\widehat{\sigma}^2_{t,k}-\sigma^2_{t,k}|\le C\sqrt{\frac{s_0\log(K\vee n)}{n}}+\frac{Cs_0}{m},\ \textit{and}$$
$$\max_{1\le t<k\le K}\left|\frac{\widehat{\sigma}_{t,k}}{\sigma_{t,k}}-1\right|\le C\sqrt{\frac{s_0\log(K\vee n)}{n}}+\frac{Cs_0}{m},$$

with probability at least $1-(K\vee n)^{-3}$.

Proof Similar to the proof of Proposition 11, we can prove that

$$\max_{1\le t<k\le K}|\widehat{\sigma}^2_{t,k}-(\Theta^{*2}_{t,k}+\Theta^*_{t,t}\Theta^*_{k,k})|\le C\sqrt{\frac{s_0\log(K\vee n)}{n}},$$

with probability at least $1-(K\vee n)^{-3}$. Then, by Lemma 23, we obtain

$$\max_{1\le t<k\le K}|\widehat{\sigma}^2_{t,k}-\sigma^2_{t,k}|\le C\sqrt{\frac{s_0\log(K\vee n)}{n}}+\frac{Cs_0}{m}.$$

The second statement can be similarly derived. ■

Having established the asymptotic normality of $\widetilde{\Theta}_{t,k}$ and the convergence rate of $\widehat{\sigma}^2_{t,k}$, we proceed with the proofs of the main results.

Proof of Theorem 4

Proof The proof follows in exactly the same manner as that of Theorem 3, but invokes Proposition 18, instead of Proposition 11, and Proposition 17, instead of Proposition 10, as one needs to establish different intermediate results, specifically tailored to estimation of the latent graph. ■

Proof of Theorem 5, Claim 2

Proof The proof follows in exactly the same manner as that of Theorem 5, Claim 1. ■

B.2. Key Lemmas for Estimators of the Latent Graph

As before, we assume in the following proofs that the event $\mathcal{E} = \{\widehat{G} = G^*\}$ holds. Using a similar argument to that used in the proof of Theorem 4, the following bounds will hold with probability at least $1 - \frac{C}{(K \vee n)^3}$.

Lemma 19 (Consistency of $\widehat{\mathbf{C}}$) *If Assumptions 1 and 2 hold, then with probability greater than* $1 - \frac{C}{(K \vee n)^3}$,

$$\|\widehat{\mathbf{C}} - \mathbf{C}^*\|_{\max} \leq C\sqrt{\frac{\log(K \vee n)}{n}},$$

for some constant C dependent only on c_1, c_2, and c_3 from Assumptions 1 and 2.

Proof Denoting $\mathbf{B}^* = \mathbf{A}^{*T}\mathbf{A}^*$, we begin by using the decomposition

$$\widehat{\mathbf{C}} - \mathbf{C}^* = \mathbf{B}^{*-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)\mathbf{A}^*\mathbf{B}^{*-1} - \mathbf{B}^{*-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*)\mathbf{A}^*\mathbf{B}^{*-1}. \tag{51}$$

Next, we can write for $i \in G_k^*$,

$$\begin{aligned}
(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*)_{i,i} &= \widehat{\gamma}_i - \gamma_i^* \\
&= \widehat{\Sigma}_{i,i} - \frac{1}{|G_k^*| - 1} \sum_{j \in G_k^*, j \neq i} \widehat{\Sigma}_{i,j} - \gamma_i^* \\
&= \widehat{\Sigma}_{i,i} - \frac{1}{|G_k^*| - 1} \sum_{j \in G_k^*, j \neq i} \widehat{\Sigma}_{i,j} - \Sigma_{i,i}^* + \frac{1}{|G_k^*| - 1} \sum_{j \in G_k^*, j \neq i} \Sigma_{i,j}^* \\
&= \widehat{\Sigma}_{i,i} - \Sigma_{i,i}^* - \frac{1}{|G_k^*| - 1} \sum_{j \in G_k^*, j \neq i} \left[\widehat{\Sigma}_{i,j} - \Sigma_{i,j}^*\right],
\end{aligned}$$

which implies that $||\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*||_{\max} \leq 2||\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*||_{\max}$. Therefore from Lemma 27 we see that

$$||\mathbf{B}^{*-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*)\mathbf{A}^*\mathbf{B}^{*-1}||_{\max} \leq \frac{2}{m}||\mathbf{B}^{*-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)\mathbf{B}^{*-1}||_{\max},$$

demonstrating that it will suffice to bound the first term in (51), which we now do.

For the first term in (51), we have

$$\begin{aligned}
&\mathbf{B}^{*-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)\mathbf{A}^*\mathbf{B}^{*-1} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}\Big\{(\mathbf{A}^*\boldsymbol{Z}_i + \boldsymbol{E}_i)(\mathbf{A}^*\boldsymbol{Z}_i + \boldsymbol{E}_i)^T - \mathbf{A}^*\mathbf{C}^*\mathbf{A}^{*T} - \mathbf{\Gamma}^*\Big\}\mathbf{A}^*\mathbf{B}^{*-1} \\
&= \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T - \mathbf{C}^* + \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{E}_i^T\mathbf{A}^*\mathbf{B}^{*-1} + \frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}\boldsymbol{E}_i\boldsymbol{Z}_i^T \\
&\quad + \frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}(\boldsymbol{E}_i\boldsymbol{E}_i^T - \mathbf{\Gamma}^*)\mathbf{A}^*\mathbf{B}^{*-1}.
\end{aligned} \tag{52}$$

Using the triangle inequality, we can apply Lemma 24, Lemma 25 and Lemma 26 to bound 52. Combining these results with that $\mathbb{P}(\mathcal{E}) \geq 1 - c_0/(d \vee n)^3$, we obtain

$$\|\widehat{\mathbf{C}} - \mathbf{C}^*\|_{\max} \leq C\sqrt{\frac{\log(K \vee n)}{n}}$$

with probability at least $1 - \frac{C}{(K \vee n)}$ for some constant C dependent only on c_1, c_2, and c_3 from Assumptions 1 and 2. ■

Lemma 20 (Concentration of Gradient and Hessian of the Loss Function) *If Assumptions 1 and 2 hold, then with probability greater than* $1 - \frac{C}{(K \vee n)^3}$*, we have that*

(a) $\max_{1 \leq k \leq K} \|\widehat{\mathbf{C}}\mathbf{\Theta}^*_{\cdot k} - \mathbf{e}_k\|_\infty \leq C_1\sqrt{\frac{\log(K \vee n)}{n}}$,

(b) $\max_{1 \leq t \leq K} \|\widehat{\mathbf{C}}_{t,-t} - \mathbf{w}_t^{*T}\widehat{\mathbf{C}}_{-t,-t}\|_\infty \leq C_2\sqrt{\frac{\log(K \vee n)}{n}}$*, and*

(c) $\max_{1 \leq t \leq K} \|\mathbf{w}_t^{*T}(\widehat{\mathbf{C}}_{-t,-t} - \mathbf{C}^*_{-t,-t})\|_\infty \leq C_2\sqrt{\frac{\log(K \vee n)}{n}}$,

for absolute constants C_1 *and* C_2 *dependent only on* c_1*,* c_2*, and* c_3 *from Assumptions 1 and 2.*

Proof Let $\mathbf{B}^* = \mathbf{A}^{*T}\mathbf{A}^*$ and observe that $\widehat{\mathbf{C}}\mathbf{\Theta}^*_{\cdot k} - \mathbf{e}_k = (\widehat{\mathbf{C}} - \mathbf{C}^*)\mathbf{\Theta}^*_{\cdot k}$. Following the decomposition (52), we can similarly show that

$$\begin{aligned}
&\mathbf{B}^{*-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)\mathbf{A}^*\mathbf{B}^{*-1}\mathbf{\Theta}^*_{\cdot k} \\
&= \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T\mathbf{\Theta}^*_{\cdot k} - \mathbf{C}^*\mathbf{\Theta}^*_{\cdot k} + \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{E}_i^T\mathbf{A}^*\mathbf{B}^{*-1}\mathbf{\Theta}^*_{\cdot k} + \frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}\boldsymbol{E}_i\boldsymbol{Z}_i^T\mathbf{\Theta}^*_{\cdot k} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}(\boldsymbol{E}_i\boldsymbol{E}_i^T - \mathbf{\Gamma}^*)\mathbf{A}^*\mathbf{B}^{*-1}\mathbf{\Theta}^*_{\cdot k}. \qquad (53)
\end{aligned}$$

As in the proof of Lemma 19, we have that

$$||\mathbf{B}^{*-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*)\mathbf{A}^*\mathbf{B}^{*-1}\mathbf{\Theta}^*_{\cdot k}||_\infty \leq \frac{2}{m}||\mathbf{B}^{*-1}\mathbf{A}^{*T}(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)\mathbf{A}^*\mathbf{B}^{*-1}\mathbf{\Theta}^*_{\cdot k}||_\infty,$$

demonstrating that again it will suffice to bound the first term in 53.

Note that $||\mathbf{\Theta}^*_{\cdot k}||_2 \leq \lambda_{\max}(\mathbf{\Theta}^*) \leq c_1^{-1}$. Therefore, by using the triangle inequality, we can apply Lemma 24, Lemma 25 and Lemma 26 to bound the first term in 53. Combining these results and $\mathbb{P}(\mathcal{E}) \geq 1 - C/(d \vee n)^3$, we obtain

$$\max_{1 \leq k \leq K} \|\widehat{\mathbf{C}}\mathbf{\Theta}^*_{\cdot k} - \mathbf{e}_k\|_\infty \leq C_1\sqrt{\frac{\log(K \vee n)}{n}},$$

with probability at least $1 - \frac{C}{(K \vee n)^3}$ for some constant C_1 dependent only on c_1, c_2, and c_3 from Assumptions 1 and 2.

For the remaining two claims, we can rewrite

$$\mathbf{w}_t^* = \left(C_{t,t}^* - \mathbf{C}_{-t,t}^{*T}(\mathbf{C}_{-t,-t}^*)^{-1}\mathbf{C}_{-t,t}^*\right)\boldsymbol{\Theta}_{-t,t}^* = \frac{1}{\Theta_{t,t}^*}\boldsymbol{\Theta}_{-t,t}^*$$

by the block matrix inverse formula. Using Lemma 30, it follows that $||\mathbf{w}_t^*||_2 \leq \lambda_{\max}(\boldsymbol{\Theta}^*)\max_t C_{t,t}^*$. Then we see that

$$\begin{aligned}\max_{1\leq t\leq K}||\widehat{\mathbf{C}}_{t,-t} - \mathbf{w}_t^{*T}\widehat{\mathbf{C}}_{-t,-t}||_\infty &= \max_{1\leq t\leq K}||(\widehat{\mathbf{C}}_{t,-t} - \mathbf{C}_{t,-t}^*) - \mathbf{w}_t^{*T}(\widehat{\mathbf{C}}_{-t,-t} - \mathbf{C}_{-t,-t}^*)||_\infty \\ &\leq \underbrace{\max_{1\leq t\leq K}||(\widehat{\mathbf{C}}_{t,-t} - \mathbf{C}_{t,-t}^*)||_\infty}_{\text{(i)}} + \underbrace{\max_{1\leq t\leq K}||\mathbf{w}_t^{*T}(\widehat{\mathbf{C}}_{-t,-t} - \mathbf{C}_{-t,-t}^*)||_\infty}_{\text{(ii)}}\end{aligned}$$

Clearly, using Lemma 19, (i) is bounded with high probability. Likewise, Lemma 19 demonstrates that $\widehat{\mathbf{C}}_{-t,i} - \mathbf{C}_{-t,i}^*$ is a sub-exponential random vector with parameters dependent only on c_1, c_2, and c_3 from Assumptions 1 and 2. Thus $\mathbf{w}_t^{*T}(\widehat{\mathbf{C}}_{-t,i} - \mathbf{C}_{-t,i}^*)$ is sub-exponential and because $||\mathbf{w}_t^*||_2 \leq \lambda_{\max}(\boldsymbol{\Theta}^*)\max_t C_{t,t}^*$, we obtain that

$$\max_{1\leq t\leq K}\|\widehat{\mathbf{C}}_{t,-t} - \mathbf{w}_t^{*T}\widehat{\mathbf{C}}_{-t,-t}\|_\infty \leq C_2\sqrt{\frac{\log(K\vee n)}{n}}$$

with probability at least $1 - \frac{C}{(K\vee n)^3}$ for some constant C_2. C_2 is dependent only on c_1, c_2, and c_3 from Assumptions 1 and 2. The final result is bounded by the previous one, concluding the proof. ■

Lemma 21 (Consistency of Initial Estimators) *If Assumptions 1 and 2 hold, then*

(a) $\max_{1\leq k\leq K}\|\widehat{\Theta}_{\cdot k} - \Theta_{\cdot k}^*\|_1 \leq C_1 s_0\sqrt{\frac{\log(K\vee n)}{n}}$,

(b) $\max_{1\leq k\leq K}\|\widehat{\Theta}_{\cdot k} - \Theta_{\cdot k}^*\|_2 \leq C_1\sqrt{\frac{s_0\log(K\vee n)}{n}}$,

(c) $\max_{1\leq t\leq K}\|\widehat{\mathbf{v}}_t - \mathbf{v}_t^*\|_1 \leq C_2 s_0\sqrt{\frac{\log(K\vee n)}{n}}$, *and*

(d) $\max_{1\leq k\leq t\leq K}|(\widehat{\mathbf{v}}_t - \mathbf{v}_t^*)^T\widehat{\mathbf{C}}(\widehat{\boldsymbol{\Theta}}_{\cdot k} - \boldsymbol{\Theta}_{\cdot k}^*)| \leq C_3\frac{s_0\log(K\vee n)}{n}$,

with probability at least $1 - \frac{C_4}{(K\vee n)^3}$. *$C_1$, C_2, C_3, C_4 are constants, dependent only upon c_0, c_1, c_2, and c_3 from Assumptions 1 and 2.*

Proof Below, the constants C_a, C_a', C_b, C_b', C_b'', C_c and C_c' will depend only upon c_0, c_1, c_2, and c_3 from Assumptions 1 and 2.

We first prove part (b). The proof of part (a) is similar. Let $\widehat{\boldsymbol{\Delta}} = \widehat{\mathbf{w}}_t - \mathbf{w}_t^*$, noting that we can consider $\mathbf{w}_t$ instead of $\mathbf{v}_t$ as the t^{th} entries in both the estimated and true value are 1. By S denote the support of $\mathbf{w}_t^*$. $\mathbf{w}_t^*$ is s_0-sparse because $\mathbf{w}_t^*$ is a multiple of $\boldsymbol{\Theta}_{-t,k}^*$, which we know to be s_0-sparse.

By Lemma 20, there exists C_b such that for $\lambda \geq C_b\sqrt{\frac{\log(K\vee n)}{n}}$, $\mathbf{w}_t^*$ is feasible for (22) with probability at least $1 - \frac{C_b'}{(K\vee n)^3}$. Assuming $\mathbf{w}_t^*$ is feasible, then it follows by definition that $||(\mathbf{w}_t^*)_S||_1 \geq ||(\widehat{\mathbf{w}}_t)_S||_1 + ||(\widehat{\mathbf{w}}_t)_{\bar{S}}||_1$. This in turn implies by the triangle inequality that $||\widehat{\mathbf{\Delta}}_S||_1 \geq ||\widehat{\mathbf{\Delta}}_{\bar{S}}||_1$. Letting $\lambda = C_b\sqrt{\frac{\log(K\vee n)}{n}}$, it follows from the triangle inequality that

$$\begin{aligned}||\widehat{\mathbf{C}}_{-t,-t}\widehat{\mathbf{\Delta}}||_\infty &\leq ||\widehat{\mathbf{w}}_t^T\widehat{\mathbf{C}}_{-t,-t} - \mathbf{C}_{t,-t}||_\infty + ||\mathbf{w}_t^{*T}\widehat{\mathbf{C}}_{-t,-t} - \mathbf{C}_{t,-t}||_\infty \\ &\leq 2C_b\sqrt{\frac{\log(K\vee n)}{n}}.\end{aligned}$$

In addition, note that $||\widehat{\mathbf{\Delta}}||_1 \leq 2||\widehat{\mathbf{\Delta}}_S||_1 \leq 2\sqrt{s_0}||\widehat{\mathbf{\Delta}}_S||_2 \leq 2\sqrt{s_0}||\widehat{\mathbf{\Delta}}||_2$. Therefore combining with the above, this gives

$$\begin{aligned}\widehat{\mathbf{\Delta}}^T\widehat{\mathbf{C}}_{-t,-t}\mathbf{\Delta} &\leq ||\widehat{\mathbf{\Delta}}||_1||\widehat{\mathbf{C}}_{-t,-t}\widehat{\mathbf{\Delta}}||_\infty \\ &\leq 2C\sqrt{\frac{\log(K\vee n)}{n}}||\widehat{\mathbf{\Delta}}||_1 \\ &\leq 4C_b\sqrt{\frac{s_0\log(K\vee n)}{n}}||\widehat{\mathbf{\Delta}}||_2.\end{aligned}$$

From Lemma 29, $\widehat{\mathbf{\Delta}}^T\widehat{\mathbf{C}}_{-t,-t}\widehat{\mathbf{\Delta}} \geq \frac{4c_1}{3}||\widehat{\mathbf{\Delta}}||_2^2$ with probability at least $1 - \frac{C_b''}{(K\vee n)^3}$. Therefore

$$\begin{aligned}||\widehat{\mathbf{\Delta}}||_2 &\leq \frac{16Cc_1}{3}\sqrt{\frac{s_0\log(K\vee n)}{n}}, \text{ and} \\ ||\widehat{\mathbf{\Delta}}||_1 &\leq \frac{32C_bs_0c_1}{3}\sqrt{\frac{\log(K\vee n)}{n}},\end{aligned}$$

with probability at least $1 - \frac{\max\{C_b', C_b''\}}{(K\vee n)^3}$.

To obtain part (c), first let $\delta = \max_{1\leq k\leq t\leq K}|(\widehat{\mathbf{v}}_t - \mathbf{v}_t^*)^T\widehat{\mathbf{C}}(\widehat{\mathbf{\Theta}}_{\cdot k} - \mathbf{\Theta}_{\cdot k}^*)|$ and then by applying Holder's inequality and the triangle inequality, we obtain

$$\begin{aligned}\delta &\leq \max_{1\leq k\leq t\leq K}||\widehat{\mathbf{v}}_t - \mathbf{v}_t^*||_1||\widehat{\mathbf{C}}(\widehat{\mathbf{\Theta}}_{\cdot k} - \mathbf{\Theta}_{\cdot k}^*)||_\infty \\ &\leq \max_{1\leq k\leq t\leq K}||\widehat{\mathbf{v}}_t - \mathbf{v}_t^*||_1\left(||\widehat{\mathbf{C}}\widehat{\mathbf{\Theta}}_{\cdot k} - \boldsymbol{e}_k||_\infty + ||\widehat{\mathbf{C}}\mathbf{\Theta}_{\cdot k}^* - \boldsymbol{e}_k||_\infty\right). \qquad (54)\end{aligned}$$

With choice of λ as above, the KKT conditions give that $||\widehat{\mathbf{C}}\widehat{\mathbf{\Theta}}_{\cdot k} - \boldsymbol{e}_k||_\infty \leq C_b\sqrt{\frac{\log(K\vee n)}{n}}$. From Lemma 20, we have that $||\widehat{\mathbf{C}}\mathbf{\Theta}_{\cdot k}^* - \boldsymbol{e}_k||_\infty \leq C\sqrt{\frac{\log(K\vee n)}{n}}$ with probability at least $1 - \frac{C}{(K\vee n)^3}$. Using part (b), we get that $\max_{1\leq t\leq K}\|\widehat{\mathbf{v}}_t - \mathbf{v}_t^*\|_1 \leq Cs_0\sqrt{\frac{\log(K\vee n)}{n}}$ with probability at least $1 - \frac{C}{(K\vee n)^3}$. The desired result now follows from (54). ■

Lemma 22 (CLT for the Pseudo-Score Function) *Recall that*

$$\sigma_{tk}^2 = \mathbb{E}(\Theta_{tt}^*\mathbf{v}_t^{*T}(\bar{\mathbf{C}}^{(i)}\mathbf{\Theta}_{\cdot k}^* - \mathbf{e}_k))^2,$$

with $\bar{\mathbf{C}}^{(i)}$ defined in (30). Let F_n denote the CDF of $n^{-1/2}\mathbf{v}_t^{*T}(\widehat{\mathbf{C}}\boldsymbol{\Theta}^*_{\cdot k} - \mathbf{e}_k)/(\sigma_{tk}/\Theta^*_{tt})$. If Assumptions 1 and 2 hold, then we have

$$\max_{1\le t<k\le K} \sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)| \le C(n^{-1/2} + (d\vee n)^{-3}).$$

where C is a constant dependent only upon c_0, c_1, and c_2.

Proof Denote by $\mathcal{E}$ the event that $\widehat{G} = G^*$. We have

$$F_n(x) - \Phi(x) \le \widetilde{F}_n(x) - \Phi(x) + \mathbb{P}(\bar{\mathcal{E}}),$$

where $\widetilde{F}_n(x)$ is the CDF of $n^{-1/2}\sum_{i=1}^n \mathbf{v}_t^{*T}(\bar{\mathbf{C}}^{(i)}\boldsymbol{\Theta}^*_{\cdot k} - \mathbf{e}_k)/(\sigma_{tk}/\Theta^*_{tt})$. To control $\widetilde{F}_n(x) - \Phi(x)$, we now verify the Lyapunov condition. As in the proof of Lemma 20, we can write

$$\mathbf{v}_t^{*T}(\bar{\mathbf{C}}^{(i)}\boldsymbol{\Theta}^*_{\cdot k} - \mathbf{e}_k) = \mathbf{v}_t^{*T}(\bar{\mathbf{C}}^{(i)} - \mathbf{C}^*)\boldsymbol{\Theta}^*_{\cdot k}.$$

From Lemmas 24 - 26 we see that the entries in $\boldsymbol{Q}_i = (\bar{\mathbf{C}}^{(i)} - \mathbf{C}^*)\boldsymbol{\Theta}^*_{\cdot k}$ are sub-exponential with parameters $\alpha = C_1$ and $\nu = C_2$ which depend only upon $\lambda_{\max}(\boldsymbol{\Theta}^*)$, $\max_k \gamma^*_k$, and $\max_t C^*_{t,t}$.

Recall the definition of $\boldsymbol{v}_t^*$: $(\boldsymbol{v}_t^*)_t = 1$ and $(\boldsymbol{v}_t^*)_{-t} = -\mathbf{w}_t^* = -(\mathbf{C}^*_{-t,-t})^{-1}\mathbf{C}^*_{-t,t}$. By the block matrix inverse formula, we can rewrite

$$\mathbf{w}_t^* = \left(C^*_{t,t} - \mathbf{C}^{*T}_{-t,t}(\mathbf{C}^*_{-t,-t})^{-1}\mathbf{C}^*_{-t,t}\right)\boldsymbol{\Theta}^*_{-t,t} = \frac{1}{\Theta^*_{t,t}}\boldsymbol{\Theta}^*_{-t,t}.$$

Using Lemma 30, it follows that $||\mathbf{w}_t^*||_2 \le \lambda_{\max}(\boldsymbol{\Theta}^*)\max_t C^*_{t,t}$ and $||\mathbf{v}_t^*||_2 \le \lambda_{\max}(\boldsymbol{\Theta}^*)\max_t C^*_{t,t} + 1$. From Lemma 33 and the above, $\mathbf{v}_t^{*T}\boldsymbol{Q}_i$ is sub-exponential with parameters $\alpha = C_1$ and $\nu = ||\mathbf{v}_t^*||_2 C_2 \le \left(\lambda_{\max}(\boldsymbol{\Theta}^*)\max_t C^*_{t,t} + 1\right)C_2$. Therefore, $\mathbf{v}_t^{*T}(\bar{\mathbf{C}}^{(i)}\boldsymbol{\Theta}^*_{\cdot k} - \mathbf{e}_k)$ has third moments bounded above by some constant ρ that depends only upon $\lambda_{\max}(\boldsymbol{\Theta}^*)$, $\max_k \gamma^*_k$, and $\max_t C^*_{t,t}$. All three quantities are bounded above by constants per Assumptions 1 and 2. Thus, $\max_{1\le t<k\le K}\sup_x(\widetilde{F}_n(x) - \Phi(x)) \le Cn^{-1/2}$ by the classical Berry-Esseen Theorem, and therefore

$$\max_{1\le t<k\le K} \sup_{x\in\mathbb{R}}(F_n(x) - \Phi(x)) \le C(n^{-1/2} + (d\vee n)^{-3}).$$

Similarly, it can be shown that $\sup_{x\in\mathbb{R}}(\Phi(x) - F_n(x)) \le C(n^{-1/2} + (d\vee n)^{-3})$. This completes the proof. ■

Lemma 23 (Approximation for Asymptotic Variance) *Under Assumptions 1 and 2, we have that*

$$\sigma^2_{tk} = \Theta^{*2}_{t,k} + \Theta^*_{t,t}\Theta^*_{k,k} + \Delta, \tag{55}$$

where $|\Delta| \le \frac{Cs_0}{m}$ and C is a constant dependent only upon c_1, c_2 and c_3.

Proof Recall that $\sigma^2_{tk} = \mathbb{E}(\Theta^*_{tt}\mathbf{v}_t^{*T}(\bar{\mathbf{C}}^{(i)}\boldsymbol{\Theta}^*_{\cdot k} - \mathbf{e}_k))^2$. Using the identity $\mathrm{vec}(\mathbf{M}_1\mathbf{M}_2\mathbf{M}_3) = (\mathbf{M}_3^T \otimes \mathbf{M}_1)^T\,\mathrm{vec}(\mathbf{M}_2)(\mathbf{M}_3^T \otimes \mathbf{M}_1)$, we have

$$\sigma^2_{tk} = (\Theta^*_{t,t})^2(\boldsymbol{\Theta}^{*T}_{\cdot k} \otimes \boldsymbol{v}_t^*)^T\mathbb{E}\left[\mathrm{vec}(\bar{\mathbf{C}}^{(i)})\,\mathrm{vec}(\bar{\mathbf{C}}^{(i)})^T\right](\boldsymbol{\Theta}^{*T}_{\cdot k} \otimes \boldsymbol{v}_t^*). \tag{56}$$

Computing the expectation: After some straightforward, albeit lengthy, algebra we can show that

$$\mathbb{E}\left[\operatorname{vec}(\bar{\mathbf{C}}^{(i)})\operatorname{vec}(\bar{\mathbf{C}}^{(i)})^T\right] = \mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3,$$

where $\mathbf{M}_1 := \mathbf{C}^* \otimes \mathbf{C}^*$ and $\mathbf{M}_2 := [\mathbf{C}^*_{\cdot j}\mathbf{C}^{*T}_{\cdot i}]_{ij}$. The matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ contribute to the first two terms in (55). The term $\mathbf{M}_3 := \mathbb{E}\left[\operatorname{vec}(\bar{\mathbf{C}}^{(i)})\operatorname{vec}(\bar{\mathbf{C}}^{(i)})^T\right] - \mathbf{M}_1 - \mathbf{M}_2$, however, is unique to the latent graph and contributes the higher order term Δ in (55).
Evaluating the first order terms: By (56), we have

$$\sigma^2_{tk} = (\Theta^*_{t,t})^2(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T(\mathbf{M}_1 + \mathbf{M}_2)(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t) + \Delta$$

with Δ defined as

$$\Delta := (\Theta^*_{t,t})^2(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T\mathbf{M}_3(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t). \tag{57}$$

Next, we observe that

$$\begin{aligned}(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T\mathbf{M}_1(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t) &= \boldsymbol{\Theta}^{*T}_{\cdot k}\mathbf{C}^*\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^{*T}_t\mathbf{C}^*\boldsymbol{v}^*_t \\ &= \frac{\Theta^*_{k,k}}{\Theta^*_{t,t}},\end{aligned}$$

where we used that $\boldsymbol{v}^{*T}_t\mathbf{C}^*\boldsymbol{v}^*_t = (\Theta^*_{t,t})^{-1}$. Similarly, we can find that

$$(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T\mathbf{M}_2(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t) = \frac{\Theta^{*2}_{tk}}{\Theta^{*2}_{tt}}.$$

Bounding the higher order terms: What remains is to bound the magnitude of the term Δ in (57). Lengthy algebra yields:

$$|(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T\mathbf{M}_3(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)| \leq \frac{C'}{m}(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T\left(\mathbf{M}_4 + \mathbf{M}_5 + \mathbf{M}_6 + \mathbf{M}_7\right)(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t),$$

where C' depends only on c_1, c_2 and c_3. Here, $\mathbf{M}_4 = \mathbf{I} \otimes (\mathbf{1}\mathbf{1}^T)$. For $l = 5, 6, 7$ the matrices $\mathbf{M}_l$ are defined block-wise by

$$\mathbf{M}_{5;ij} := \begin{cases}\mathbf{1}\mathbf{e}_i^T & \text{if } i \neq j \\ \mathbf{0} & \text{o/w}\end{cases} \text{ and } \mathbf{M}_{6;ij} := \begin{cases}\mathbf{e}_j\mathbf{1}^T & \text{if } i \neq j \\ \mathbf{0} & \text{o/w}\end{cases} \text{ and } \mathbf{M}_{7;ij} := \begin{cases}\mathbf{I} & \text{if } i \neq j \\ \mathbf{0} & \text{o/w}\end{cases}.$$

Further lengthy algebra gives:

$$|(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T\mathbf{M}_4(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)| \leq s_0\frac{2}{c_1^2}(1 + \frac{c_2^2}{c_1^2}),$$

$$|(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T\mathbf{M}_5(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)| \leq s_0\frac{2\sqrt{2}}{c_1^2}(1 + \frac{c_2^2}{c_1^2}),$$

$$|(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)^T\mathbf{M}_6(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}^*_t)| \leq s_0\frac{2\sqrt{2}(c_1^2 + c_2^2)}{c_1^4}\sqrt{1 + \frac{c_2^2}{c_1^2}},$$

and

$$|(\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}_t^*)^T \mathbf{M}_7 (\boldsymbol{\Theta}^*_{\cdot k} \otimes \boldsymbol{v}_t^*)| \leq s_0 \frac{2(c_1^2 + c_2^2)}{c_1^4}.$$

Plugging these bounds into the expression for Δ in (57), we obtain

$$|\Delta| \leq \frac{C s_0}{m},$$

concluding the proof. ■

Appendix C. Concentration Results

The lemmas below provide important results regarding the concentration properties of some of the estimators $\widehat{\mathbf{C}}$ and variables $\boldsymbol{Z}$.

Lemma 24

(a) $\boldsymbol{Z}_i \boldsymbol{Z}_i^T$ *consists of entries which are sub-exponential with parameters* $\alpha = 4 \max_t (C^*_{t,t})^2$ *and* $\nu = 2\sqrt{2} \max_t (C^*_{t,t})^2$,

(b) $\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^T - \mathbf{C}^* \right\|_{\max} \geq C \max_t (C^*_{t,t})^2 \sqrt{\frac{\log(K \vee n)}{n}} \right) \leq \frac{2}{(K \vee n)^3}$

(c) $\mathbf{Z}_i \mathbf{Z}_i^T \mathbf{u}$ *consists of entries which are sub-exponential with parameters* $\alpha = 4 \max_t (C^*_{t,t})^2$ *and* $\nu = 2\sqrt{2} ||\mathbf{u}||_2 \max_t (C^*_{t,t})^2$, *and*

(d) $\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n (\boldsymbol{Z}_i \boldsymbol{Z}_i^T - \mathbf{C}^*) \mathbf{u} \right\|_{\max} \geq C ||\mathbf{u}||_2 \max_t (C^*_{t,t})^2 \sqrt{\frac{\log(K \vee n)}{n}} \right) \leq \frac{2}{(K \vee n)^3}$,

where $C = 4\sqrt{3}$ *and* $\mathbf{u} \in \mathbb{R}^K$.

Proof From Lemma 31, each element in the matrices $\boldsymbol{Z}_i \boldsymbol{Z}_i^T$ are sub-exponential with parameters $\alpha = 4 \max_t (C^*_{t,t})^2$ and $\nu = 2\sqrt{2} \max_t (C^*_{t,t})^2$. Therefore by Corollary 33, the entries in $\frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^T$ are sub-exponential with parameters $\alpha = \frac{4}{n} \max_t (C^*_{t,t})^2$ and $\nu = \frac{2\sqrt{2}}{\sqrt{n}} \max_t (C^*_{t,t})^2$. Therefore by the tail bound for sub-exponential random variables, we see that

$$\begin{aligned}
&\mathbb{P}\left(\left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^T - \mathbf{C}^* \right)_{s,t} \geq D_1 \sqrt{\frac{\log(K \vee n)}{n}} \right) \\
&\leq \begin{cases} \exp\left(-\frac{(\log(K \vee n)) D_1^2}{16 \max_t (C^*_{t,t})^4} \right) & \text{if } 0 \leq D_1 \sqrt{\frac{\log(K \vee n)}{n}} \leq 2 \max_t (C^*_{t,t})^2 \\ \exp\left(\frac{-D_1 \sqrt{n \log(K \vee n)}}{8 \max_t (C^*_{t,t})^2} \right) & \text{if } D_1 \sqrt{\frac{\log(K \vee n)}{n}} > 2 \max_t (C^*_{t,t})^2 \end{cases}
\end{aligned}$$

for arbitrary $D_1 > 0$. Observe that for n sufficiently large, $D_1 \sqrt{\log(K \vee n)/n} \leq 2 \max_t (C^*_{t,t})^2$, and thus we need only consider this case. Choose $D_1 \geq 4\sqrt{3} \max_t (C^*_{t,t})^2$. Then it is clear that

$$\mathbb{P}\left(\left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^T - \mathbf{C}^* \right)_{s,t} \geq D_1 \sqrt{\frac{\log(K \vee n)}{n}} \right) \leq \frac{1}{(K \vee n)^5}.$$

By applying the union bound across all entries in the matrix, we get the desired result that

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{Z}_i^T-\mathbf{C}^*\right\|_{\max}\geq D_1\sqrt{\frac{\log(K\vee n)}{n}}\right)\leq\frac{2}{(K\vee n)^3},$$

concluding the proof of parts (a) and (b). The proof of (c) and (d) are very similar and omitted. ■

Lemma 25 *Let* $C=2\sqrt{3}$, $\sigma_s^2=\frac{1}{|G_s^*|^2}\sum_{i\in G_s^*}\gamma_i$, *and* $\boldsymbol{M}_i=\boldsymbol{Z}_i\boldsymbol{E}_i^T\mathbf{A}^*\mathbf{B}^{*-1}$. *Then,*

(a) $(\boldsymbol{M}_i)_{s,t}$ *is sub-exponential with parameters* $\alpha=\sqrt{2}\max\left(\sigma_s^2,C_{t,t}^*\right)$ *and* $\nu=\sqrt{2}\max\left(\sigma_s^2,C_{t,t}^*\right)$,

(b) $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{M}_i\right\|_{\max}\geq C\max_t(\sigma_t^2\vee C_{t,t}^*)\sqrt{\frac{\log(K\vee n)}{n}}\right)\leq\frac{2}{(K\vee n)^3}$

(c) $(\boldsymbol{M}_i\mathbf{u})_s$ *is sub-exponential with parameters* $\alpha=\sqrt{2}\max\left(\sigma_s^2,C_{t,t}^*\right)$ *and* $\nu=\sqrt{2}||\mathbf{u}||_2\max\left(\sigma_s^2,C_{t,t}^*\right)$, *and*

(d) $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{M}_i\mathbf{u}\right\|_{\infty}\geq C\max_t(\sigma_t^2\vee C_{t,t}^*)||\mathbf{u}||_2\sqrt{\frac{\log(K\vee n)}{n}}\right)\leq\frac{2}{(K\vee n)^3}$,

where $\mathbf{u}\in\mathbb{R}^K$.

Proof Let $\boldsymbol{M}=\sum_{i=1}^{n}\boldsymbol{M}_i$. From Lemma 27, $\boldsymbol{Y}_1=\mathbf{B}^{*-1}\mathbf{A}^{*T}\boldsymbol{E}_1$ is a K-dimensional vector where the k^{th} entry is given by

$$(Y_1)_k=\frac{1}{|G_k^*|}\sum_{i\in G_k^*}(E_1)_i.$$

Because the errors are all independent mean zero Gaussian random variables, $(Y_1)_s\sim\mathcal{N}(0,\sigma_s^2)$. Therefore, as Y_1 is independent of Z_1 by definition, $\mathbb{E}[(Y_1)_s(Z_1)_t]=\mathbb{E}[(Y_1)_s]\mathbb{E}[(Z_1)_t]=0$. Further, Lemma 31 gives that $(Y_1)_s(Z_1)_t$ is sub-exponential with parameters $\alpha=\nu=\sqrt{2}\max\left(\sigma_s^2,C_{t,t}^*\right)$.

Using the independence of the samples, Corollary 33 gives that $\boldsymbol{M}_{s,t}$ is sub-exponential with parameters $\alpha=\sqrt{2}\max\left(\sigma_s^2,C_{t,t}^*\right)$ and $\nu=\sqrt{2n}\max\left(\sigma_s^2,C_{t,t}^*\right)$. Then, Corollary 35 gives that for arbitrary choice of $D_1>0$,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\boldsymbol{M}_{s,t}\geq D_1\sqrt{\frac{\log(K\vee n)}{n}}\right)\\&\leq\begin{cases}\exp\left(-\frac{(\log(K\vee n))D_1^2}{4\max\left(\sigma_s^2,C_{t,t}^*\right)^2}\right) & \text{if } 0\leq D_1\sqrt{\frac{\log(K\vee n)}{n}}\leq\sqrt{2}\max\left(\sigma_s^2,C_{t,t}^*\right)\\ \exp\left(\frac{-D_1\sqrt{n\log(K\vee n)}}{\sqrt{2}\max\left(\sigma_s^2,C_{t,t}^*\right)}\right) & \text{if } D_1\sqrt{\frac{\log(K\vee n)}{n}}>\sqrt{2}\max\left(\sigma_s^2,C_{t,t}^*\right).\end{cases}\end{aligned}$$

Observe that for n sufficiently large, $D_1\sqrt{\log K/n} \leq \sqrt{2}\max\left(\sigma_s^2, C_{t,t}^*\right)$. If we choose $D_1 \geq 2\sqrt{3}\max\left(\sigma_s^2, C_{t,t}^*\right)$, then we obtain that for n sufficiently large,

$$\mathbb{P}\left(\frac{1}{n}\boldsymbol{M}_{s,t} \geq D_1\sqrt{\frac{\log(K\vee n)}{n}}\right) \leq \frac{1}{(K\vee n)^5}.$$

Then by the union bound we can obtain

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{E}_i^T\mathbf{A}^*\mathbf{B}^{*-1}\right\|_{\max} \geq D_1\sqrt{\frac{\log(K\vee n)}{n}}\right) \leq \frac{2}{(K\vee n)^3}$$

for $D_1 \geq 2\sqrt{3}\max\left(\max_s \sigma_s^2, \max_t C_{t,t}^*\right)$, concluding the proof of parts (a) and (b). The proof of (c) and (d) are very similar and omitted. ■

Lemma 26 *Recall that* $m = \min_k |G_k^*|$, *and let* $\boldsymbol{M}_i = \mathbf{B}^{*-1}\mathbf{A}^{*T}\boldsymbol{E}_i\boldsymbol{E}_i^T\mathbf{A}^*\mathbf{B}^{*-1}$ *and* $\boldsymbol{\mu} = \mathbf{B}^{*-1}\mathbf{A}^{*T}\boldsymbol{\Gamma}^*\mathbf{A}^*\mathbf{B}^{*-1}$. *Then,*

(a) $(\boldsymbol{M}_i - \boldsymbol{\mu})_{t,k}$ *is sub-exponential with parameters* $\alpha_{t,k} = \frac{\sqrt{2}}{|G_k^*||G_t^*|}\max_{i\in G_t^*\cup G_k^*}\gamma_i^*$ *and* $\nu_{t,k} = \sqrt{\frac{2}{|G_k^*||G_t^*|}}\max_{i\in G_t^*\cup G_k^*}\gamma_i^*$,

(b) $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{M}_i - \boldsymbol{\mu}\right\|_{\max} \geq C\max_k\gamma_k\sqrt{\frac{\log(K\vee n)}{nm^2}}\right) \leq \frac{2}{(K\vee n)^3}$

(c) $((\boldsymbol{M}_i - \boldsymbol{\mu})\,\mathbf{u})_t$ *is sub-exponential with parameters* $\alpha_t = \frac{\sqrt{2}}{m^2}\max_k\gamma_k^*$ *and* $\nu_t = \sqrt{\frac{2}{m^2}}||\mathbf{u}||_2\max_k\gamma_k^*$, *and*

(d) $\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n(\boldsymbol{M}_i - \boldsymbol{\mu})\,\mathbf{u}\right\|_{\max} \geq C||\mathbf{u}||_2\max_k\gamma_k\sqrt{\frac{\log(K\vee n)}{nm^2}}\right) \leq \frac{2}{(K\vee n)^3}$,

where $C = 2\sqrt{3}$ *and and* $\mathbf{u} \in \mathbb{R}^K$.

Proof To obtain the results, we first bound the sum $\boldsymbol{M} := \sum_{i=1}^n \boldsymbol{M}_i$ entrywise. From Lemma 28, Corollary 32 and Corollary 33, we have that $(\boldsymbol{M}_i)_{t,k}$ is sub-exponential with parameters

$$\alpha_{t,k} = \frac{\sqrt{2}}{|G_k^*||G_t^*|}\max_{i\in G_t^*\cup G_k^*}\gamma_i^* \quad \text{and} \quad \nu_{t,k} = \sqrt{\frac{2}{|G_k^*||G_t^*|}}\max_{i\in G_t^*\cup G_k^*}\gamma_i^*.$$

Therefore $\frac{1}{n}M_{t,k}$ is sub-exponential with parameters

$$\alpha = \frac{\sqrt{2}}{n|G_k^*||G_t^*|}\max_{i\in G_t^*\cup G_k^*}\gamma_i^* \quad \text{and} \quad \nu = \sqrt{\frac{2}{n|G_k^*||G_t^*|}}\max_{i\in G_t^*\cup G_k^*}\gamma_i^*.$$

Next, observe that $\mu_{t,k} = \mathbb{E}[M_{t,k}]$ and denote $N = \max_{i \in G_t^* \cup G_k^*} \gamma_i^*$. Then, from Lemma 34, we obtain

$$\mathbb{P}\left(\frac{1}{n}M_{t,k} - \mu_{t,k} \geq D_1\sqrt{\frac{\log(K \vee n)}{n|G_k^*||G_t^*|}}\right)$$
$$\leq \begin{cases} \exp\left(-\frac{(\log(K\vee n))D_1^2}{4N^2}\right) & \text{if } 0 \leq D_1\sqrt{\frac{\log(K\vee n)}{n|G_k^*||G_t^*|}} \leq \sqrt{2}N \\ \exp\left(\frac{-D_1\sqrt{n\log(K\vee n)}}{\sqrt{2}N}\right) & \text{if } D_1\sqrt{\frac{\log(K\vee n)}{n|G_k^*||G_t^*|}} > \sqrt{2}N. \end{cases}$$

Observe that for n sufficiently large, $D_1\sqrt{\frac{\log(K\vee n)}{n|G_k^*||G_t^*|}} \leq \sqrt{2}N$. If we choose $D_1 \geq 2\sqrt{3}N$, then we obtain that for n sufficiently large,

$$\mathbb{P}\left(\frac{1}{n}M_{t,k} - \mu_{t,k} \geq D_1\sqrt{\frac{\log(K \vee n)}{n|G_k^*||G_t^*|}}\right) \leq \frac{1}{(K \vee n)^5}.$$

Therefore by taking the union bound, lower bounding $n|G_k^*||G_t^*|$ by nm^2 and choosing $D_1 \geq 2\sqrt{3}\max_k \gamma_k^*$,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{B}^{*-1}\mathbf{A}^{*T}(\boldsymbol{E}_i\boldsymbol{E}_i^T - \boldsymbol{\Gamma}^*)\mathbf{A}^*\mathbf{B}^{*-1}\right\|_{\max} \geq D_1\sqrt{\frac{\log(K \vee n)}{nm^2}}\right) \leq \frac{2}{(K \vee n)^3},$$

concluding the proof of parts (a) and (b). The proof of (c) and (d) are very similar and omitted. ■

Appendix D. Auxiliary Technical lemmas

Lemma 27 *For $1 \leq k \leq K$, denote $m_k = |G_k^*|$. Then the matrix $(\mathbf{A}^{*T}\mathbf{A}^*)^{-1}\mathbf{A}^{*T}$ is a $K \times d$ dimensional matrix given as*

$$[(\mathbf{A}^{*T}\mathbf{A}^*)^{-1}\mathbf{A}^{*T}]_{k,i} = \begin{cases} \frac{1}{m_k} & \text{if } i \in G_k^* \\ 0 & \text{otherwise.} \end{cases}$$

Proof First, we must calculate $\mathbf{B}^{*-1}\mathbf{A}^{*T}$. For $1 \leq k \leq K$, denote $m_k = |G_k^*|$ and let $\mathbf{e}_k$ be a unit vector in $\mathbb{R}^K$ with 1 on the k position and 0 otherwise. Without loss of generality, we permute the rows of $\mathbf{A}^*$ such that for any $1 \leq k \leq K$ $\mathbf{A}_{j\cdot}^* = \mathbf{e}_k$, for $\sum_{i=1}^{k-1} m_i + 1 \leq j < \sum_{i=1}^k m_i + 1$ – that is rows are ordered according to ascending group index. Here, for notational simplicity, we let $m_0 = 0$. Thus, $\mathbf{A}^{*T}\mathbf{A}^* = \text{diag}(m_1, ..., m_K)$ and the result follows immediately. ■

Lemma 28 *The matrices $\mathbf{B}^{*-1}\mathbf{A}^{*T}(\boldsymbol{E}_i\boldsymbol{E}_i^T)\mathbf{A}^*\mathbf{B}^{*-1}$ and $\mathbf{B}^{*-1}\mathbf{A}^{*T}(\Gamma^*)\mathbf{A}^*\mathbf{B}^{*-1}$ are given by*

$$(\mathbf{B}^{*-1}\mathbf{A}^{*T}(\boldsymbol{E}_i\boldsymbol{E}_i^T)\mathbf{A}^*\mathbf{B}^{*-1})_{t,k} = \frac{1}{|G_t^*G_k^*|}\sum_{p\in G_t^*}\sum_{q\in G_k^*} E_{i,p}E_{i,q}$$

and

$$(\mathbf{B}^{*-1}\mathbf{A}^{*T}(\Gamma^*)\mathbf{A}^*\mathbf{B}^{*-1})_{t,k} = \begin{cases} \frac{1}{|G_t^*|^2}\sum_{p\in G_t^*}\gamma_p^* & \text{if } t = k \\ 0 & \text{otherwise.} \end{cases}$$

Proof The result can be obtained by a straightforward computation. ■

Lemma 29 (Restricted Eigenvalue Condition for $\widehat{\mathbf{C}}$) *If Assumptions 1 and 2 hold, then the matrix $\widehat{\mathbf{C}}$ satisfies with probability at least $1 - \frac{C}{(K\vee n)^3}$,*

$$\kappa \leq \min\left\{\frac{\mathbf{v}^T\widehat{\mathbf{C}}\mathbf{v}}{||\mathbf{v}||_2^2} : \mathbf{v}\in\mathbb{R}^K\setminus\{0\}, ||\mathbf{v}_{\bar{S}}||_1 \leq 3||\mathbf{v}_S||_1\right\}, \text{ and}$$

$$\kappa \leq \min\left\{\frac{\mathbf{v}^T\widehat{\mathbf{C}}_{-t,-t}\mathbf{v}}{||\mathbf{v}||_2^2} : \mathbf{v}\in\mathbb{R}^K\setminus\{0\}, ||\mathbf{v}_{\bar{S}'}||_1 \leq 3||\mathbf{v}_{S'}||_1\right\},$$

where $\kappa \geq \frac{3}{4c_1} > 0$.

Proof We begin by proving the first claim. By Lemma 19, we have that $\|\widehat{\mathbf{C}} - \mathbf{C}^*\|_{\max} \leq C_1\sqrt{\frac{\log(K\vee n)}{n}}$ with high probability. Therefore, for K sufficiently large and for any $\mathbf{v}\in\mathbb{R}^K\setminus\{0\}$,

$$\frac{\mathbf{v}^T\widehat{\mathbf{C}}\mathbf{v}}{||\mathbf{v}||_2^2} \geq \frac{3}{4}\frac{\mathbf{v}^T\mathbf{C}^*\mathbf{v}}{||\mathbf{v}||_2^2}.$$

The proof is then done for $\kappa = \frac{3}{4c_1}$ as we assume the minimum eigenvalue of $\mathbf{C}^*$ is bounded below by c_0^{-1}. The proof of the second claim is identical because $\mathbf{C}^*$ is positive semidefinite, and it is well known that the minimum eigenvalue of any principal submatrix $\mathbf{C}^*_{-t,-t}$ is bounded below by $\lambda_{\min}(\mathbf{C}^*) \geq c_0^{-1}$. ■

Lemma 30 *Let $\mathbf{M}$ be a $n\times n$ positive definite matrix and denote its inverse by $\mathbf{L}$. Then, for all $i = 1,\ldots,n$*

$$M_{i,i}L_{i,i} \geq 1.$$

Proof By the block matrix inverse formula, it follows that

$$M_{i,i}^{-1} = L_{i,i} - \mathbf{L}_{-i,i}^T\mathbf{L}_{-i,-i}^{-1}\mathbf{L}_{-i,i}. \tag{58}$$

Because $\mathbf{M}$ is positive definite, so is $\mathbf{L}$. Recall that a matrix is positive definite if and only if all its principal minors are also positive definite. Therefore, $\mathbf{L}_{-i,-i}$ is positive definite, as is $\mathbf{L}_{-i,-i}^{-1}$. Therefore, $\mathbf{L}_{-i,i}^T\mathbf{L}_{-i,-i}^{-1}\mathbf{L}_{-i,i} \geq 0$ and (58) becomes $M_{i,i}^{-1} \leq L_{i,i}$. Lastly, if a matrix is positive definite, all its diagonal elements must be nonnegative, giving that $M_{i,i}L_{i,i} \geq 1$ as desired. ■

Appendix E. Basic Tail Bounds for Random Variables

This section collects some basic tail probability results for random variables. The proof is standard and omitted.

Lemma 31 *Let $\mathbf{Y} = (Y_1, Y_2)$ be a jointly Gaussian random vector with covariance matrix $\mathbf{C}$. Then $Y_1 Y_2$ is sub-exponential with parameters $\alpha = 4\lambda_{\max}(\mathbf{C}_Y)$ and $\nu = 2\sqrt{2}\lambda_{\max}(\mathbf{C}_Y)$.*

Corollary 32 *Let $Y_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(0, \sigma_2^2)$ where $\sigma_1^2 \geq \sigma_2^2$. Then $Y_1 Y_2$ is sub-exponential with parameters $\alpha = \sqrt{2}\sigma_1^2$ and $\nu = \sqrt{2}\sigma_1^2$.*

Corollary 33 *Consider $\sum_{i=1}^n X_i$ where X_i are centered, independent sub-exponential random variables. Then $Y = \sum_{i=1}^n X_i$ is sub-exponential with parameters $\alpha = \max_i \alpha_i$ and $\nu = \sqrt{\sum_{i=1}^n \nu_i^2}$.*

Lemma 34 (Tail Bound for Sub-Exponential Random Variables) *Let X be a sub-exponential random variable with mean μ and parameters α and ν. Then*

$$\mathbb{P}(X - \mu \geq t) \leq \begin{cases} \exp(-\frac{t^2}{2\nu^2}) & \text{for } 0 \leq t \leq \frac{\nu^2}{\alpha} \\ \exp(-\frac{t}{2\alpha}) & \text{for } t > \frac{\nu^2}{\alpha}. \end{cases}$$

Corollary 35 *Consider $Y = \sum_{i=1}^n X_i$, where X_i are centered, independent sub-exponential random variables. Let $\alpha = \max_i \alpha_i$ and $\nu = \sqrt{\sum_{i=1}^n \nu_i^2}$. Then,*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i \geq t) \leq \begin{cases} \exp(-\frac{nt^2}{2\nu^2/n}) & \text{for } 0 \leq t \leq \frac{\nu^2}{n\alpha} \\ \exp(-\frac{nt}{2\alpha}) & \text{for } t > \frac{\nu^2}{n\alpha}. \end{cases}$$

Appendix F. Construction of a Pre-Clustering Variance Estimator

We include in this section the construction of the pre-clustering estimator of Γ needed as an input of the PECOK algorithm of Section 2.2 above. For any $a, b \in [d]$, define

$$V(a,b) := \max_{c,d \in [p]\setminus\{a,b\}} \frac{\left|(\widehat{\Sigma}_{ac} - \widehat{\Sigma}_{ad}) - (\widehat{\Sigma}_{bc} - \widehat{\Sigma}_{bd})\right|}{\sqrt{\widehat{\Sigma}_{cc} + \widehat{\Sigma}_{dd} - 2\widehat{\Sigma}_{cd}}}, \tag{59}$$

with the convention $0/0 = 0$. Guided by the block structure of Σ, we define

$$b_1(a) := \operatorname*{argmin}_{b \in [p]\setminus\{a\}} V(a,b) \quad \text{and} \quad b_2(a) := \operatorname*{argmin}_{b \in [p]\setminus\{a, b_1(a)\}} V(a,b),$$

to be two elements "close" to a, that is two indices $b_1 = b_1(a)$ and $b_2 = b_2(a)$ such that the empirical covariance difference $\widehat{\Sigma}_{b_i c} - \widehat{\Sigma}_{b_i d}$, $i = 1, 2$, is most similar to $\widehat{\Sigma}_{ac} - \widehat{\Sigma}_{ad}$, for all variables c and d not equal to a or b_i, $i = 1, 2$. It is expected that $b_1(a)$ and $b_2(a)$ either belong to the same group as a, or belong to some "close" groups. Then, our estimator $\widetilde{\Gamma}$ is a diagonal matrix, defined by

$$\widetilde{\Gamma}_{aa} = \widehat{\Sigma}_{aa} + \widehat{\Sigma}_{b_1(a)b_2(a)} - \widehat{\Sigma}_{ab_1(a)} - \widehat{\Sigma}_{ab_2(a)}, \quad \text{for } a = 1, \ldots, d. \tag{60}$$

Intuitively, $\widetilde{\Gamma}_{aa}$ should be close to $\Sigma_{aa} + \Sigma_{b_1(a)b_2(a)} - \Sigma_{ab_1(a)} - \Sigma_{ab_2(a)}$, which is equal to Γ_{aa} in the favorable event where both $b_1(a)$ and $b_2(a)$ belong to the same group as a.

In general, $b_1(a)$ and $b_2(a)$ cannot be guaranteed to belong to the same group as a. Nevertheless, these two surrogates $b_1(a)$ and $b_2(a)$ are close enough to a so that $\|\widetilde{\Gamma} - \Gamma\|_{\max} \lesssim |\Gamma|_{\max}\sqrt{\log d/n}$. This last fact and the above construction are shown in Bunea et al. (2018).

Appendix G. Comparison with Cai et al (2016)

In our work, we bound $\lambda_{\min}(\boldsymbol{S}^*) \geq c_1$ and $\max_t S^*_{t,t} \leq c_2$ and the sparsity of $\Omega_{\cdot k}$ by s_1. We show that the CLIME estimator satisfies

$$\|\widehat{\boldsymbol{\Omega}}_{\cdot k} - \boldsymbol{\Omega}^*_{\cdot k}\|_1 \lesssim s_1\sqrt{\frac{\log(K \vee n)}{n}}. \tag{61}$$

Let us denote our parameter space by

$$\mathcal{G}_1 = \{\boldsymbol{\Omega} : \max_k \|\Omega_{\cdot k}\|_0 \leq s_1, \max_t\{(\boldsymbol{\Omega}^{-1})_{tt}\} \leq c_2, \lambda_{\max}(\boldsymbol{\Omega}) \leq 1/c_1\}.$$

By contrast, Cai et al. (2016) considered the following parameter space for the precision matrix $\boldsymbol{\Omega} = \boldsymbol{S}^{-1}$ (in the exact sparse case)

$$\mathcal{G} = \{\boldsymbol{\Omega} : \max_k \|\Omega_{\cdot k}\|_0 \leq s_1, \|\boldsymbol{\Omega}\|_1 \leq M_n, \kappa(\boldsymbol{\Omega}) = \lambda_{\max}(\boldsymbol{\Omega})/\lambda_{\min}(\boldsymbol{\Omega}) \leq M_1\},$$

where $\|\boldsymbol{\Omega}\|_1$ is the matrix ℓ_1-norm, M_n is allowed to increase with n and M_1 is a constant bound for the condition number. Their minimax lower bound over $\mathcal{G}$ depends on M_n, that is for any $\widehat{\boldsymbol{\Omega}}_{\cdot k}$,

$$\sup_{\boldsymbol{\Omega}\in\mathcal{G}} \|\widehat{\boldsymbol{\Omega}}_{\cdot k} - \boldsymbol{\Omega}_{\cdot k}\|_1 \gtrsim M_n s_1\sqrt{\frac{\log K}{n}}. \tag{62}$$

It seems that the upper bound (61) and the lower bound (62) contradict with each other. However, this is not the case because the parameter space $\mathcal{G}_1$ under which the estimator is developed is different from the parameter space $\mathcal{G}$ in the lower bound. This therefore opens up the possibility of obtaining different rates over different parameter spaces, for instance one which does not depend on $||\boldsymbol{\Omega}^*||_1$, like in our case. However, neither parameter space can be viewed as a full relaxation of the other. Below we give explicit examples that show how the parameter spaces differ.

Consider the sequence of precision matrices, indexed by K

$$\boldsymbol{\Omega}_K = |\mathcal{S}(K)|\mathbf{I} + \mathbf{1}_{\mathcal{S}(K)}\mathbf{1}^\top_{\mathcal{S}(K)} \in \mathbb{R}^K,$$

where $\mathcal{S}(K)$ is an arbitrary subset of $[K] = \{1, 2, ..., K\}$ with $|\mathcal{S}(K)| = K^{1/4}$ and $\mathbf{1}_{\mathcal{S}(K)}$ is the vector with 1's in the indices indicated by $\mathcal{S}(K)$ and 0's elsewhere. By using the Sherman-Morrison formula, we can verify that the corresponding covariance matrix is

$$\boldsymbol{S}_K = \frac{1}{|\mathcal{S}(K)|}\mathbf{I} - \frac{1}{2|\mathcal{S}(K)|^2}\mathbf{1}_{\mathcal{S}(K)}\mathbf{1}^\top_{\mathcal{S}(K)}.$$

It is easy to check that

$$\lambda_{\min}(\mathbf{\Omega}_K) = |\mathcal{S}(K)| = K^{1/4},$$
$$\lambda_{\max}(\mathbf{\Omega}_K) = 2|\mathcal{S}(K)| = 2K^{1/4},$$
$$\lambda_{\min}(\boldsymbol{S}_K) = \frac{1}{2|\mathcal{S}(K)|} = K^{-1/4}/2.$$

Thus $\kappa(\mathbf{\Omega}_K) = 2$, and therefore Theorem 4.2 from Cai et al. (2016) can be applied to this sequence of parameters $\mathbf{\Omega}_K$ with $s_1 = K^{1/4} + 1$ and $M_n = 2K^{1/4}$. However, $\mathbf{\Omega}_K$ does not belong to our parameter space $\mathcal{G}_1$ because for any choice of c_1, once K is sufficiently large, $\lambda_{\min}(\boldsymbol{S}_K) < c_1$.

Similarly, we can modify the above example such that $\mathbf{\Omega}_K$ satisfies our condition but does not belong to $\mathcal{G}$. For instance, consider

$$\mathbf{\Omega}_K = \mathbf{I} - \frac{1}{1 + |\mathcal{S}(K)|}\mathbf{1}_{\mathcal{S}(K)}\mathbf{1}_{\mathcal{S}(K)}^\top \in \mathbb{R}^K.$$

We can show that

$$\boldsymbol{S}_K = \mathbf{I} + \mathbf{1}_{\mathcal{S}(K)}\mathbf{1}_{\mathcal{S}(K)}^\top,$$

and

$$\lambda_{\min}(\mathbf{\Omega}_K) = \frac{1}{K^{1/4} + 1},$$
$$\lambda_{\max}(\mathbf{\Omega}_K) = 1,$$
$$\lambda_{\min}(\boldsymbol{S}_K) = 1.$$

Our conditions $\lambda_{\min}(\boldsymbol{S}_K) \geq c_1$ and $\max_t(\boldsymbol{S}_K)_{t,t} \leq c_2$ hold, and thus we obtain the rate (61) for the CLIME estimator. However, $\kappa(\mathbf{\Omega}_K) = K^{1/4} + 1 \to \infty$, as $K \to \infty$. Thus, $\mathbf{\Omega}_K$ does not belong to the parameter space $\mathcal{G}$, and Theorem 4.2 in Cai et al. (2016) is not applicable.

Since our goal is to make inference on each entry of the precision matrix say Ω_{tk}, a lower bound on the minimum eigenvalue on $\boldsymbol{S}$ is a mild and standard assumption for any high-dimensional inference on graphical models. After some careful analysis as explained in the main text, we show that under Assumption 4.1 and 4.2 the upper bound for the CLIME estimator does not depend on $||\mathbf{\Omega}^*||_1$ or $\lambda_{\max}(\boldsymbol{S}^*)$.

Appendix H. B-H procedures for FDR control

In this section, we analyze the theoretical properties of the B-H procedure for the cluster-average graph. The procedure for latent variable graph is identical. Recall that the test statistic for $H_{0,tk} : \Omega^*_{t,k} = 0$ is $\widetilde{W}_{t,k}$. The B-H procedure is defined as follows. Given the desired FDR level α, define

$$\widehat{\rho} = \inf\Big\{0 \leq \rho \leq 2\sqrt{\log K} : \frac{2(1 - \Phi(\rho))|\mathcal{H}|}{\max\{\sum_{1\leq t\leq k\leq K} I(|\widetilde{W}_{t,k}| \geq \rho), 1\}} \leq \alpha\Big\},$$

where $|\mathcal{H}| = K(K-1)/2$ is the total number of hypotheses. If $\widehat{\rho}$ does not exist, then we set $\widehat{\rho} = 2\sqrt{\log K}$. In the above definition, we restrict $\rho \leq 2\sqrt{\log K}$ in order to apply the

Cramer-type moderate deviation result (Liu, 2013). The B-H procedure says that we reject $H_{0,tk}$ if $|\widetilde{W}_{t,k}| \geq \widehat{\rho}$. The FDR and FDP are defined as

$$\text{FDP} = \frac{\sum_{(t,k)\in\mathcal{H}_0} I(|\widetilde{W}_{t,k}| \geq \widehat{\rho})}{\max\{\sum_{(t,k)\in\mathcal{H}} I(|\widetilde{W}_{t,k}| \geq \widehat{\rho}), 1\}},$$

where $\mathcal{H}_0 \coloneqq \{(t,k) : 1 \leq t < k \leq K, \text{ such that } \Omega^*_{t,k} = 0\}$, and $\text{FDR} = \mathbb{E}(\text{FDP})$. Let

$$\mathcal{A} = \left\{(t,k) : \frac{\Omega^*_{t,k}}{\sqrt{\Omega^*_{k,k}\Omega^*_{t,t}}} \geq 4\sqrt{\log K/n}\right\}$$

denote the set of strong signals. The following theorem on the B-H procedure holds.

Proposition 36 *Assume that the conditions in Theorem 3 hold. Let $K \leq n^r$ for some $r > 0$. In addition, assume that $|\mathcal{A}| \geq \sqrt{\log\log K}$, $s_1 \log^{3/2}(K \vee n)/n^{1/2} = o(1)$ and $s_1 = O(K^c)$ for some $c < 1/2$. Then, as $n, K \to \infty$, we have*

$$\frac{\text{FDP}}{|\mathcal{H}_0|/|\mathcal{H}|} = \alpha + o_p(1) \quad \textit{and} \quad \frac{\text{FDR}}{|\mathcal{H}_0|/|\mathcal{H}|} = \alpha + o(1).$$

The proof of this proposition follows from Theorem 3.1 in Liu (2013). The conditions $|\mathcal{A}| \geq \sqrt{\log\log K}$ and $s_1 = O(K^c)$ are equation (12) in Liu (2013). The condition $s_1 \log^{3/2}(K \vee n)/n^{1/2} = o(1)$ together with Theorem 3 guarantees that

$$\max_{1\leq t<k\leq K} \sup_{x\in\mathbb{R}} \left|\mathbb{P}(\widehat{T}_{t,k} \leq x) - \Phi(x)\right| = o(1/\sqrt{\log(K \vee n)}), \tag{63}$$

which is equivalent to condition (13) in Liu (2013). We refer to Liu (2013) for the detailed proof. Proposition 36 implies that the B-H procedure can control FDR asymptotically under certain assumption. However, our numerical results show that this procedure may fail to control FDR in our model when K is relatively large. One possible reason is that the $o_p(1)$ or $o(1)$ terms in the approximation of FDP or FDR in the above proposition are not sufficiently small in simulations due to the dependence of the test statistics. Thus, if the goal of the data analysis is to find statistically reliable scientific discoveries, the B-Y method, while quite conservative, may be the one that's more suitable for this purpose.

References

Jessica R Andrews-Hanna, Jay S Reidler, Jorge Sepulcre, Renee Poulin, and Randy L Buckner. Functional-anatomic fractionation of the brain's default network. *Neuron*, 65 (4):550–562, 2010.

Rina Foygel Barber and Mladen Kolar. Rocket: Robust confidence intervals via kendall's tau for transelliptical graphical models, 2015.

Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S. Margulies, and R. Cameron Craddock. The Neuro Bureau ADHD-200 Preprocessed repository. *NeuroImage*, 2015. ISSN 10959572. doi: 10.1016/j.neuroimage.2016.06.034.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

Peter J Bickel. One-step huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434, 1975.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Y Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

Florentina Bunea. Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics*, pages 898–927, 2004.

Florentina Bunea, Christophe Giraud, Xi Luo, Martin Royer, and Nicolas Verzelen. Model assisted variable clustering: minimax-optimal recovery and algorithms, 2018.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

T Tony Cai, Weidong Liu, and Harrison Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2):455–488, 2016.

Carson Eisenach and Han Liu. Efficient, certifiably optimal clustering with applications to latent variable graphical models. *Mathematical Programming*, 2019. doi: 10.1007/s10107-019-01375-2.

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):405–421, 2017.

Ethan X Fang, Yang Ning, and Han Liu. Testing and confidence intervals for high dimensional proportional hazards models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1415–1437, 2017.

Huijie Feng and Yang Ning. High-dimensional mixed graphical model with ordinal data: Parameter estimation and statistical inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 654–663, 2019.

Michael D Fox, Abraham Z Snyder, Justin L Vincent, Maurizio Corbetta, David C Van Essen, and Marcus E Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102 (27):9673–9678, 2005.

Peter Fransson. Spontaneous low-frequency bold signal fluctuations: An fmri investigation of the resting-state default mode of brain function hypothesis. *Human brain mapping*, 26 (1):15–29, 2005.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Christophe Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

Quanquan Gu, Yuan Cao, Yang Ning, and Han Liu. Local and global inference for high dimensional gaussian copula graphical models, 2015.

Jana Jankova and Sara van de Geer. Confidence intervals for high-dimensional inverse covariance estimation, 2014.

Jana Janková and Sara van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *Test*, 26(1):143–162, 2017.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression, 2013.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37:42–54, 2009.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact inference after model selection via the lasso, 2013.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Weidong Liu. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, pages 2948–2978, 2013.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

Xi Luo. A hierarchical graphical model for big inverse covariance estimation with an application to fmri. *arXiv preprint arXiv:1403.4698*, 2014.

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

Matey Neykov, Yang Ning, Jun S Liu, and Han Liu. A unified theory of confidence regions and testing for high dimensional estimating equations. *Statistical Science*, 2018.

Yang Ning and Han Liu. High-dimensional semiparametric bigraphical models. *Biometrika*, 100(3):655–670, 2013.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195, 2017.

Yang Ning, Tianqi Zhao, Han Liu, et al. A likelihood ratio framework for high-dimensional semiparametric regression. *The Annals of Statistics*, 45(6):2299–2327, 2017.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, and Steven E Petersen. Functional network organization of the human brain. *Neuron*, 17(724):665–678, 2011. doi: 10.1016/j.neuron.2011.09.006.

Marcus E Raichle. The brain's default mode network. *Annual review of neuroscience*, 38: 433–447, 2015.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model, 2013.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Christopher G Small, Jinfang Wang, and Zejiang Yang. Eliminating multiple root problems in estimation. *Statistical Science*, 15(4):313–341, 2000.

Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045, 2009.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. Technical report, 2012.

Kean Ming Tan, Yang Ning, Daniela M Witten, and Han Liu. Replicates in high dimensions, with applications to latent variable graphical models. *Biometrika*, 103(4):761–777, 2016.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures, 2014.

Sara van de Geer, Peter Bühlmann, and Ya'acov Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models, 2013.

A. V. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 1998.

Nicolas Verzelen. *Gaussian graphical models and Model selection*. PhD thesis, Université Paris Sud-Paris XI, 2008.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

Zhuoran Yang, Yang Ning, and Han Liu. On semiparametric exponential family graphical models. *The Journal of Machine Learning Research*, 19(1):2314–2372, 2018.

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(8):2261–2286, 2010.

M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. ISSN 1467-9868. doi: 10.1111/rssb.12026.

Yuan Zhou, Karl J Friston, Peter Zeidman, Jie Chen, Shu Li, and Adeel Razi. The hierarchical organization of the default, dorsal attention and salience networks in adolescents and young adults. *Cerebral cortex*, 28(2):726–737, 2017.