Commun. Math. Res.
doi: 10.4208/cmr.2020-0001

Vol. **36**, No. 1, pp. 68-92
March 2020

Precision Matrix Estimation by Inverse Principal Orthogonal Decomposition

Cheng Yong Tang[1,*], Yingying Fan[2] and Yinfei Kong[3]

[1] *Department of Statistical Science, Fox School of Business, Temple University, Philadelphia, PA 19122, USA.*

[2] *Marshall School of Business, University of Southern California, Los Angeles, CA 90089, USA.*

[3] *Department of Information Systems and Decision Sciences, Mihaylo College of Business and Economics, California State University, Fullerton, CA 92831, USA.*

Received 3 January 2020; Accepted 11 February 2020

Abstract. We investigate the structure of a large precision matrix in Gaussian graphical models by decomposing it into a low rank component and a remainder part with sparse precision matrix. Based on the decomposition, we propose to estimate the large precision matrix by inverting a principal orthogonal decomposition (IPOD). The IPOD approach has appealing practical interpretations in conditional graphical models given the low rank component, and it connects to Gaussian graphical models with latent variables. Specifically, we show that the low rank component in the decomposition of the large precision matrix can be viewed as the contribution from the latent variables in a Gaussian graphical model. Compared with existing approaches for latent variable graphical models, the IPOD is conveniently feasible in practice where only inverting a low-dimensional matrix is required. To identify the number of latent variables, which is an objective of its own interest, we investigate and justify an approach by examining the ratios of adjacent eigenvalues of the sample covariance matrix. Theoretical properties, numerical examples, and a real data application demonstrate the merits of the IPOD approach in its convenience, performance, and interpretability.

Key words: High-dimensional data analysis, latent Gaussian graphical model, precision matrix.

[*] Corresponding author. *Email addresses:* `yongtang@temple.edu` (C. Y. Tang), `fanyingy@marshall.usc.edu` (Y. Fan), `yikong@fullerton.edu` (Y. Kong)

1 Introduction

Exploring how subjects and/or variables are connected to each other in various systems is one of the most common and important problems in practical applications. Examples of such investigations are regularly seen in scenarios including regression analysis, Gaussian graphical models, classification, principal component analysis and many more. Investigations of this kind are encountered even more often in practical applications in recent popular areas such as finance, biological and medical studies, meteorological and astronomical research, among others. Because of the general interest on the connections between individuals, the scale of these investigations can easily grow beyond a practical and manageable scope — for example, considering the complexity of possible associations among human genes. Therefore, parsimonious modeling approaches are critically important for generating practical, feasible, and interpretable statistical analyses when exploring the association structures of the target systems in many contemporary studies.

For studying the connections between subjects/variables, precision matrix, the inverse of a covariance matrix, is a crucial device in many statistical analyses including Gaussian graphical models [12], discriminant analysis, dimension reduction, and investment portfolio analysis. There has been an increasing interest in penalized likelihood approaches for estimating large precision matrices in recent literature; see, for example, [7,8,10,13,15–17,19] and references therein. In Gaussian graphical models, the precision matrix has the interpretation that each of its zero elements implies the conditional independence of the corresponding pair of individuals given the information from all other individuals. In the corresponding graph consisting of a vertex set and an edge set, such conditional independence means that there is no edge between the corresponding pair of vertices representing the individuals.

With latent variables, analyzing Gaussian graphical models becomes substantially more difficult; see [4] in which a penalized likelihood approach is investigated. More specifically, the interpretation of the graphical model becomes less clear if the impact of latent variables is not incorporated in the large precision matrix. Additionally, the unknown number of the latent variables also poses new challenges, both computationally in optimizing the penalized likelihood function and practically in developing most appropriate interpretations of the graphical models. A remarkable feature of the Gaussian graphical model with latent variables is that although the underlying the true precision matrix is sparse indicating small number of connected vertices in the corresponding graph, latent variables generally cause a non-sparse observable precision matrix of the variables excluding those latent ones. Because a fundamental assumption of many existing penalized likelihood based methods for Gaussian graphical models is that the underlying true precision matrix is sparse, they are expected to fail to consistently estimating the precision matrix without incorporating the latent variables. Moreover, non-sparse precision matrices between observed variables are often seen in data collected in many problems in finance, biomedical studies, gene-environment associations, and so forth.

In this study, we demonstrate that the precision matrix of the observable variables in a Gaussian graphical model with latent variables can be decomposed into two components — a low rank matrix associated with the latent variables and the remainder independent of latent variables. With this device, we show that if it is a sparse large precision matrix is associated with the combined observable and latent variables, then the precision matrix associated with the remainder component is sparse. This device also enable us to develop a new approach for estimating the large precision matrix of the observable variables by inverting a principal orthogonal decomposition (IPOD) of the covariance matrix that disentangling these two components. More specifically, the contribution from the latent variables is captured by a low rank matrix, which can be effectively recovered from data by using the principal component analysis (PCA), a popular approach in factor analysis. After removing the impact due to the latent variables, we show that the sparse precision matrix of the remainder part can be consistently estimated by applying the constrained l_1 minimization (CLIME) method of [3]. Moreover, we observe that the large precision matrix of the observed variables, though being non-sparse, depends on the remainder component only through its precision matrix. Hence, upon obtaining an estimation by using our IPOD approach, only inverting a small matrix is required to estimate the large and non-sparse precision matrix of observable variables.

The number of unknown latent variables is unknown *a priori* in practice, leading to a challenging problem of its own importance. To identify it, we examine the ratios between adjacent ordered eigenvalues of the sample covariance matrix of the observable variables. We show that the maximum of the ratios is effective for estimating the number of the latent variables, which is also the rank of the low rank component in the aforementioned decomposition. As an independent interest of its own, our method for identifying the number of latent variables is also useful for identifying the number of factors in a factor model with high data dimensionality.

Our investigation contributes to the area of large precision matrix estimation in the following two aspects. First, our IPOD approach for estimating large Gaussian graphical models with latent variables is convenient and adaptive. When there is no latent variable, our approach reduces to the estimation of sparse Gaussian graphical models. With few latent variables, the IPOD approach provides a useful structural device for parsimoniously modeling and practically interpreting a large precision matrix. From a practical perspective, such a benefit can provide additional insights for statistical analyses of real data when solving practical problems. Two concrete examples are elaborated in Section 3. By applying the IPOD approach for investigating the dynamic structure of stock returns, we are able to reveal some interesting and sensible association structure of the idiosyncratic component that may not be adequately explained by a systematic component of the factor model. In another application, for example, exploring the associations of university webpage content, reasonable and interpretable structure can also be detected by the IPOD approach even after removing the systematic component. Second, our theoretical analysis also reveals some appealing properties of the IPOD approach. Our theory shows that the IPOD approach enjoys similar asymptotic properties as the POET approach in [6]. More specifically, the estimation error of the IPOD approach is shown to converge to zero under both the Frobenius norm and spectra norm, as the dimensionality of precision matrix and the sample size go to infinity. In addition, the impact of unobservable factors on the estimation error vanishes as the dimensionality of the precision matrix diverges. In the absence of latent variables, our IPOD approach reduces to the CLIME approach in [3], and the corresponding estimation error bounds coincide with the ones therein.

The rest of this paper is organized as follows. In Section 2 we present our model setting, the proposed approach, and the main theoretical results. In Section 3, we first use two real-life data examples to demonstrate the appealing performance of the IPOD approach and then conduct extensive simulation studies to compare our approach with some existing ones. Finally, technical conditions are summarized in Appendix A and proofs are outlined in Appendix B.

2 The IPOD approach and main results

2.1 Model setting and method

Let us consider a sequence of multivariate random vectors $(\mathbf{y}_t', \mathbf{f}_t')'$ $(t=1,\cdots,T)$, where $\mathbf{y}_t \in \mathbb{R}^p$ is observable and $\mathbf{f}_t \in \mathbb{R}^K$ collects unobservable variables. For a Gaussian graphical model, we assume that for each $t=1,\cdots,T$,

$$(\mathbf{y}_t', \mathbf{f}_t')' \sim N(\mathbf{0}, \boldsymbol{\Theta}^{-1}), \tag{2.1}$$

where $\mathbf{\Theta}$ is the precision matrix of size $p+K$. Here, we take the mean vector as $\mathbf{0}$ without loss of generality. The above framework includes multivariate time series with t standing for time, and general multivariate models with t representing the index of independent observations. Write

$$\mathbf{\Theta}=\begin{pmatrix}\mathbf{\Omega}_\varepsilon & \mathbf{\Phi}\\ \mathbf{\Phi}' & \mathbf{\Omega}_f\end{pmatrix},$$

where $\mathbf{\Omega}_\varepsilon$ is of size $p\times p$, and $\mathbf{\Phi}$ and $\mathbf{\Omega}_f$ are of appropriate sizes. Conditioning on $\mathbf{f}_t$, the observable vector $\mathbf{y}_t$ has the distribution $\mathbf{y}_t|\mathbf{f}_t\sim N(-\mathbf{\Omega}_\varepsilon^{-1}\mathbf{\Phi}\mathbf{f}_t,\mathbf{\Omega}_\varepsilon^{-1})$, which suggests the following regression model for $\mathbf{y}_t$:

$$\mathbf{y}_t=\mathbf{B}\mathbf{f}_t+\boldsymbol{\varepsilon}_t, \tag{2.2}$$

where $\mathbf{B}=-\mathbf{\Omega}_\varepsilon^{-1}\mathbf{\Phi}\in\mathbb{R}^{p\times K}$, and $\boldsymbol{\varepsilon}_t\sim N(\mathbf{0},\mathbf{\Omega}_\varepsilon^{-1})$ is independent of $\mathbf{f}_t$. When $\mathbf{\Omega}_\varepsilon$ is diagonal, (2.2) becomes a classical factor model with loading matrix $\mathbf{B}$ and factor score $\mathbf{f}_t$. Using matrix notation, (2.2) becomes

$$\mathbf{Y}=\mathbf{F}\mathbf{B}'+\mathbf{E}, \tag{2.3}$$

where $\mathbf{Y}=(\mathbf{y}_1,\cdots,\mathbf{y}_T)'$, $\mathbf{F}=(\mathbf{f}_1,\cdots,\mathbf{f}_T)'$ and $\mathbf{E}=(\boldsymbol{\varepsilon}_1,\cdots,\boldsymbol{\varepsilon}_T)'$.

We aim at estimating $\mathbf{\Omega}$, the precision matrix of $\mathbf{y}_t$, via exploiting model (2.1) with latent component $\mathbf{f}_t$. From (2.2), it is seen that the covariance matrix

$$\mathbf{\Sigma}=\mathrm{cov}(\mathbf{y}_t)=\mathbf{B}\mathrm{cov}(\mathbf{f}_t)\mathbf{B}'+\mathbf{\Omega}_\varepsilon^{-1}. \tag{2.4}$$

Since the matrix $\mathbf{B}$ and the latent component $\mathbf{f}_t$ are generally not identifiable without extra constraints, we make the same normalization assumption as in [6]: $\mathbf{\Omega}_f=I_K$ and $\mathbf{B}'\mathbf{B}$ is diagonal, where I_K is the K-dimensional identity matrix. Hence, (2.4) is simplified as

$$\mathbf{\Sigma}=\mathbf{B}\mathbf{B}'+\mathbf{\Omega}_\varepsilon^{-1}. \tag{2.5}$$

By (2.5) and the Sherman-Morrison-Woodbury formula [9],

$$\mathbf{\Omega}=\mathbf{\Sigma}^{-1}=\mathbf{\Omega}_\varepsilon-\mathbf{\Omega}_\varepsilon\mathbf{B}(I_K+\mathbf{B}'\mathbf{\Omega}_\varepsilon\mathbf{B})^{-1}\mathbf{B}'\mathbf{\Omega}_\varepsilon. \tag{2.6}$$

Therefore the large precision matrix $\mathbf{\Omega}$ is decomposed as the sum of a sparse matrix $\mathbf{\Omega}_\varepsilon$ and a low rank matrix, and it depends on $\boldsymbol{\varepsilon}$ only through its precision matrix $\mathbf{\Omega}_\varepsilon$. We note that estimating $\mathbf{\Omega}_\varepsilon$ is of its own interest, by observing that it is the block in $\mathbf{\Theta}$ corresponding to $\mathbf{y}_t$.

We now introduce some notations to ease the future presentation. For a matrix $\mathbf{A}=(a_{ij})\in\mathbb{R}^{p\times q}$, the elementwise ℓ_1-norm is $|\mathbf{A}|_1=\sum_{i=1}^p\sum_{j=1}^q|a_{ij}|$, the elementwise ℓ_∞-norm is $|\mathbf{A}|_\infty=\max_{1\le i\le p,1\le j\le q}|a_{ij}|$, the matrix spectral norm is

$\|\mathbf{A}\|_2=\sup_{\|\mathbf{x}\|_2=1}\|\mathbf{Ax}\|_2$, the matrix 1-norm is $\|\mathbf{A}\|_1=\max_{1\le j\le q}\sum_{i=1}^{p}|a_{ij}|$, and the Frobenius norm is $\|\mathbf{A}\|_F=\sqrt{\sum_{ij}a_{ij}^2}$.

We assume that $\mathbf{\Omega}_\varepsilon=(\omega_{ij})$ is sparse and belongs to the following class of matrices:

$$\mathcal{U}=\mathcal{U}(q,s_p)=\left\{\mathbf{\Omega}_\varepsilon:\ \mathbf{\Omega}_\varepsilon\ge 0,\ \|\mathbf{\Omega}_\varepsilon\|_1\le M,\ \max_{1\le i\le p}\sum_{j=1}^{p}|\omega_{ij}|^q\le s_p\right\}, \tag{2.7}$$

where M is a positive constant, $0\le q<1$, s_p is a positive sequence depending only on p, and $\mathbf{\Omega}_\varepsilon\ge 0$ means that $\mathbf{\Omega}_\varepsilon$ is positive semidefinite. Specification (2.7) characterizes a family of sparse precision matrices. Similar specifications for precision matrices can be found in [3] and analogous specifications for sparse covariance matrices are considered in [2] and [6] among others.

We propose a two-step method that we call Inverse Principle Orthogonal Decomposition (IPOD) for estimating the precision matrix $\mathbf{\Omega}$ in (2.6). Specifically, in the first step we estimate the matrices $\mathbf{B}$ and $\mathbf{F}$ in (2.3). This is done by solving a regression problem with the least squares:

$$\min_{\mathbf{B},\mathbf{F}}\|\mathbf{Y}-\mathbf{FB}'\|_2 \ \text{ subject to } \ \mathbf{F}'\mathbf{F}=I_K \ \text{ and } \ \mathbf{B}'\mathbf{B} \ \text{ is diagonal.} \tag{2.8}$$

Denote by $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{F}}$ the resulting estimators. It was shown in [6] that the columns of $\widehat{\mathbf{F}}$ are the eigenvectors corresponding to the K largest eigenvalues of matrix $T^{-1}\mathbf{Y}'\mathbf{Y}$, and $\widehat{\mathbf{B}}=T^{-1}\mathbf{Y}'\widehat{\mathbf{F}}$. That is, $\widehat{\mathbf{B}}\widehat{\mathbf{B}}'=\sum_{i=1}^{K}\widehat{\lambda}_i\widehat{\boldsymbol{\xi}}_i\widehat{\boldsymbol{\xi}}_i'$, where $\widehat{\lambda}_1\ge\widehat{\lambda}_2\ge\cdots\ge\widehat{\lambda}_K$ are the K largest eigenvalues and $\widehat{\boldsymbol{\xi}}_1,\cdots,\widehat{\boldsymbol{\xi}}_K$ are the corresponding eigenvectors of the matrix $T^{-1}\mathbf{Y}'\mathbf{Y}$.

We now elaborate on how to specify K, the rank of $\mathbf{B}$. We propose to estimate K by examining the ratios of adjacent sample eigenvalues, that is,

$$\widehat{K}=\operatorname{argmax}_{1\le j\le (T-1)}\widehat{\lambda}_j/\widehat{\lambda}_{j+1}. \tag{2.9}$$

The validity of the ratio based method for identifying K will be justified in the next subsection. Similar method has been used in [11] to identify the number of factors K in the lower dimensional setting of $p<T$.

In the second step, we focus on the estimation of the precision matrix $\mathbf{\Omega}_\varepsilon$. Upon estimating $\mathbf{B}$ with $\widehat{\mathbf{B}}$, the covariance matrix $\mathbf{\Sigma}_\varepsilon=\mathbf{\Omega}_\varepsilon^{-1}$ can be estimated as

$$\widehat{\mathbf{\Sigma}}_\varepsilon=\widehat{\mathbf{\Sigma}}-\widehat{\mathbf{B}}\widehat{\mathbf{B}}'=\widehat{\mathbf{\Sigma}}-\sum_{i=1}^{K}\widehat{\lambda}_i\widehat{\boldsymbol{\xi}}_i\widehat{\boldsymbol{\xi}}_i'. \tag{2.10}$$

Due to high dimensionality and singularity of $\widehat{\mathbf{\Sigma}}_\varepsilon$, it is not feasible to estimate $\mathbf{\Omega}_\varepsilon$ by directly inverting $\widehat{\mathbf{\Sigma}}_\varepsilon$. In addition, even if $\widehat{\mathbf{\Sigma}}_\varepsilon$ is nonsingular, the inverse $\widehat{\mathbf{\Sigma}}_\varepsilon^{-1}$ is generally nonsparse and thus may perform poorly as an estimator of $\mathbf{\Omega}_\varepsilon$. To obtain a sparse estimator of $\mathbf{\Omega}_\varepsilon$, we propose to use the constrained l_1 minimization (CLIME) method in [3]:

$$\min_{\mathbf{\Omega}_\varepsilon \in \mathbb{R}^{p\times p}} |\mathbf{\Omega}_\varepsilon|_1 \ \text{ subject to } \ |\widehat{\mathbf{\Sigma}}_\varepsilon \mathbf{\Omega}_\varepsilon - I_p|_\infty \leq \tilde{\lambda}_n, \tag{2.11}$$

where $\tilde{\lambda}_n > 0$ is the regularization parameter. We denote by $\widehat{\mathbf{\Omega}}_\varepsilon$ the resulting estimator of the precision matrix. By substituting $\widehat{\mathbf{\Omega}}_\varepsilon$ and $\widehat{\mathbf{B}}$ into (2.6), we obtain the estimator $\widehat{\mathbf{\Omega}}$ for $\mathbf{\Omega}$.

2.2 Main results

We first justify the validity of $\hat{K}$ defined in (2.9) as an estimate of K. The proposition below shows that if there exists a gap in the ratios of adjacent population eigenvalues, then correspondingly, there is a gap in the ratios of adjacent sample eigenvalues.

Proposition 2.1. *Under Conditions* 1-3 *in Appendix A, if* $\log p = o(T^{\gamma/(2-\gamma)})$ *with* $\gamma = \gamma_1/(1+\gamma_1)$*, where* γ_1 *is defined in Condition B.*1*, then with probability at least* $1-\mathcal{O}(T^{-c_1})$*, we have*

$$1 \leq \hat{\lambda}_j / \hat{\lambda}_{j+1} \leq c_2, \quad \textit{for } j = 1,\cdots,K-1, \tag{2.12}$$

$$\hat{\lambda}_K / \hat{\lambda}_{j+1} \geq c_3 \sqrt{T/(\log p)} \to \infty, \quad \textit{for } j = K,\cdots,T-1, \tag{2.13}$$

where c_1*,* c_2 *and* c_3 *are some positive constants.*

Proposition 2.1 ensures that with probability at least $1-\mathcal{O}(T^{-c_1})$, the first $K-1$ ratios $\hat{\lambda}_j/\hat{\lambda}_{j+1}$ are bounded from above by some constant independent of p, while the Kth ratio $\hat{\lambda}_K/\hat{\lambda}_{K+1}$ diverges to infinity as $p\to\infty$. Thus, by examining the ratios of adjacent sample eigenvalues, we are able to identify the number of latent variables K consistently. In fact, in light of Proposition 2.1, the value of K can be identified as the index j where the first sudden increment in $\hat{\lambda}_j/\hat{\lambda}_{j+1}$ is observed. With some additional assumption such as that p/T is bounded away from both zero and infinity, it has been proved in [18] that the eigenvalue ratio method (2.9) can consistently estimate K. Since we are interested in the higher dimensional setting of $p \gg T$, we provide a new theoretical result below showing that (2.9) can also consistently estimate K even when p increases exponentially with T, if the observations are independently observed across t.

Theorem 2.1. *Assume that conditions of Proposition* 2.1 *hold and additionally that* $(\mathbf{y}_t',\mathbf{f}_t')'$ *with* $t=1,\cdots,T$ *are independently observed, then as* $T\to\infty$,

$$P(\widehat{K}=K)\geq 1-\mathcal{O}(T^{-c_4}),$$

where c_4 *is some positive constant.*

We next present a lemma on the properties of the covariance matrix estimator defined in (2.10) for the idiosyncratic component.

Lemma 2.1. *Assume* $\max\{(\log p)^{6/\gamma-1},K^4(\log(pT))^2\}=o(T)$ *and* $T^{1/4}K^3=o(\sqrt{p})$ *with* γ *defined in Proposition* 2.1. *Under Conditions* 1-3 *in Appendix A, we have* $|\widehat{\mathbf{\Sigma}}_\varepsilon-\mathbf{\Sigma}_\varepsilon|_\infty=O_p\big(\delta_T\big)$ *where*

$$\delta_T=\frac{K^3\sqrt{\log K}+K\sqrt{\log p}+K^2}{\sqrt{T}}+\frac{K^3}{\sqrt{p}}+\sqrt{\frac{\log p}{T}}.$$

The properties of the precision matrix estimator obtained by (2.11) is summarized in the following theorem.

Theorem 2.2. *Assume that* $\mathbf{\Omega}_\varepsilon\in\mathcal{U}(q,s_p)$ *and conditions in Lemma* 2.1 *are satisfied, if* $\tilde{\lambda}_n\geq\|\mathbf{\Omega}_\varepsilon\|_1|\widehat{\mathbf{\Sigma}}_\varepsilon-\mathbf{\Sigma}_\varepsilon|_\infty$, *then*

$$\begin{aligned}&|\widehat{\mathbf{\Omega}}_\varepsilon-\mathbf{\Omega}_\varepsilon|_\infty\leq O_p(\delta_T),\quad \|\widehat{\mathbf{\Omega}}_\varepsilon-\mathbf{\Omega}_\varepsilon\|_2=s_pO_p(\delta_T^{1-q}),\\&p^{-1}\|\widehat{\mathbf{\Omega}}_\varepsilon-\mathbf{\Omega}_\varepsilon\|_F^2\leq s_pO_p\big(\delta_T^{2-q}\big),\end{aligned}$$

where δ_T *is given in Lemma* 2.1.

It is seen that δ_T determines the convergence rates of the estimation error under various losses. The terms involving K in the definition of δ_T reflect the estimation error caused by estimating the latent variables $\mathbf{f}_t$.

The following theorem presents the asymptotic properties of the IPOD estimator $\widehat{\mathbf{\Omega}}$ for the large precision matrix $\mathbf{\Omega}$.

Theorem 2.3. *Under the assumptions of Theorem* 2.2, *we have*

$$\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}\|_2=O_p(s_q\delta_T^{1-q}),\tag{2.14}$$

$$p^{-1}\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}\|_F^2=O_p\Big(s_p\delta_T^{2-q}+\frac{Ks_p^2}{p}\delta_T^{2-2q}+\frac{K^5}{p^2}+\frac{\log p+K^3}{pT}\Big),\tag{2.15}$$

where δ_T *is the same as that in Theorem* 2.2.

When model (2.2) involves no latent variables — i.e., $K=0$ — we have $\delta_T=\sqrt{(\log p)/T}$, and the results in Theorem 2.3 become the following:

$$\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}\|_2=O_p\Big(s_q\Big(\frac{\log p}{T}\Big)^{(1-q)/2}\Big),\quad p^{-1}\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}\|_F^2=O_p\Big(s_p\Big(\frac{\log p}{T}\Big)^{(2-q)/2}\Big),$$

which are consistent with the results in [3]. We also see that as the dimensionality p of the precision matrix diverges, the terms involving K on the right-hand sides of (2.14) and (2.15) converge to zero, which means that the impact of latent variables on the estimation errors vanishes.

3 Numerical examples

3.1 University webpages

In this example, we consider the data set collected in 1997 from the "World Wide Knowledge Base" project at Carnegie Mellon University. The full data set is available from the Machine Learning Repository at the University of California, Irvine. We consider the same subset of the data as studied in [8]. The data set includes Webpages from computer science departments at Cornell University, the University of Texas, the University of Washington, and the University of Wisconsin. In our study, we use the data set of the four largest categories – student, faculty, course and project – with 544, 374, 310, and 168 Webpages, respectively. The data after some standard pre-processing are available at `http://web.ist.utl.pt/~acardoso/datasets/`.

The log-entropy weighting method was used to calculate the so-called term-document matrix $\mathbf{X}=(x_{ij})_{n\times p}$ where n and p represent the number of Webpages and the number of distinct words, respectively. Here for $i=1,\cdots,n$ and $j=1,\cdots,p$, $x_{ij}=e_j\log(1+f_{ij})$ where f_{ij} is the number of times that the jth term appears in the ith webpage, $e_j=1+\sum_{i=1}^n p_{ij}\log(p_{ij})$ is the log-entropy weight, and $p_{ij}=f_{ij}/\sum_{i=1}^n f_{ij}$. Each column of $\mathbf{X}$ is normalized to have unit ℓ_2 norm.

We apply our method to this data set for the $n=1,396$ Webpages in the four largest categories, where $p=200$ terms with the highest log-entropy weights are considered. We also assume that ε_t follows the normal distribution in this example. By pooling the data of the four categories together without distinguishing their characteristics, we estimate $K=2$ as the number of factors. The two factors explain in total 18.1% variability of the sample covariance matrix. This portion can be understood as that of the underlying common features of the Webpages, which are substantial but not dominating. By examining the loadings of the first

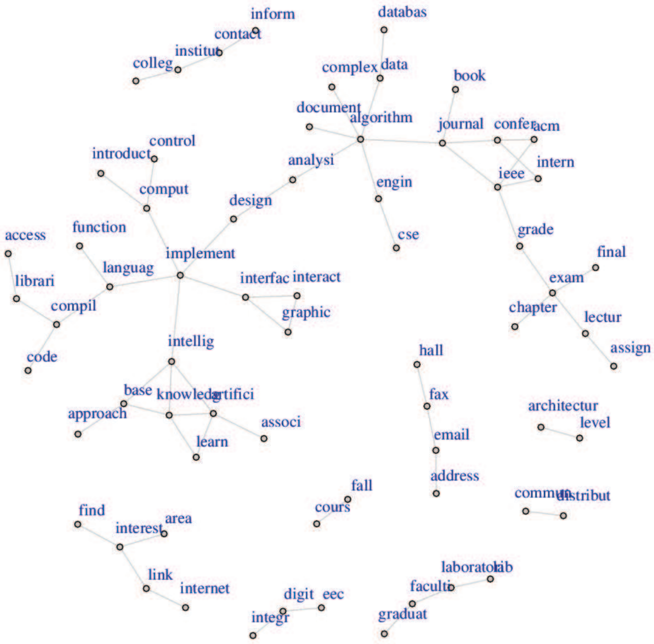


Figure 1: Graph indicating the conditional dependence structures of the Webpage example for all categories.

factor, we can see that large loadings happen for words like "system," "develop," "applic," "comput," and so forth, suggesting the computer science nature of the Webpages, and we may also understand these terms as common among all Webpages. The second factor is much more interesting as shown by the observation of large loadings with different signs on two kinds of terms. The first kind of term includes "research," "univers," "intern," "confer," "ieee," "workshop," "symposium," "proceed," "journal," and so forth, which clearly shows associations of a research nature; while the second kind includes "solut," "due," "homework," "instructor," "class," "final," "hour," "exam," "grade," "assign," and so forth, which clearly indicates features of courses Webpages. Therefore the second factor can be interpreted as the contrast between the teaching and research components of all Webpages. We also examine the possibility of using more factors, but no clear interpretations of the factors are observed.

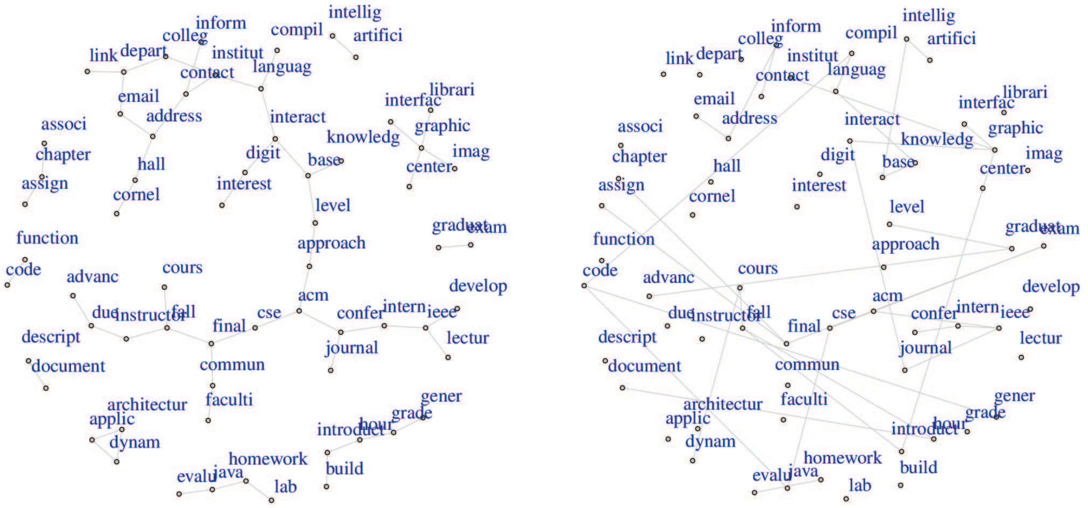
We then estimate the precision matrix $\mathbf{\Omega}_\varepsilon$. The estimated precision matrix is very sparse with only 616 nonzeros. For clarity in presenting the conditional dependence structure, we only plot the graph corresponding to the 100 most popular terms with the highest log-entropy weights in Fig. 1. First, after removing the common factors, the precision matrix of the idiosyncratic component takes a much simpler structure than that found in [8]. Second, some high-degree nodes can be identified such as “exam,” “algorithm,” and many segments are disconnected. All connected segments seem sensible.

We then examine the structure of Webpages within the student and faculty categories. For the student category with 544 Webpages, (2.9) suggests $K=3$ factors. After examining the loading matrix, we find that the first factor displays a similar pattern to the first factor identified without distinguishing the categories. Interestingly, the second factor has large values of loadings on two kinds of terms, with the first kind similar to the one found without distinguishing the categories. But the second kind has large loadings on terms such as “world,” “site,” “page,” “internet,” “web,” and so forth. Because the course Webpages belong to the teaching category and are not included here, we can see the difference between the second factor and the one found when all categories are pooled together. It may be interpreted as the contrast between the research component and other Webpage resources. The third factor has large values of loadings on terms like “techniqu,” “interfac,” “orient,” “level,” “object,” and may be understood as programming-related component.

For the faculty category, the eigenvalues also suggest $K=3$ factors. The first factor is again very similar to previous cases with a strong leaning toward computer science. In addition, the second factor has large loadings on two kinds of terms with the first kind similar to that of the student category, while the second kind being “servic,” “implement,” “memori,” “support,” “perform,” “high,” “oper,” “share,” and so forth, showing a substantial difference from that of the student category. The comparison of the graphical features of the idiosyncratic components of the student and faculty categories are given in Fig. 2. From there we can observe some similar features such as a few common high-degree nodes such as “graphic,” “final,” and “address.” However, the conditional correlation structures are seen to be different between the two categories.

3.2 Simulations

We conduct simulation studies to demonstrate the performance of the proposed approach. We first generate data from model (2.2) with $p=100$ and $T=500$ and 1000 respectively. We use $K=3$ in the data-generating scheme, and components in the loading matrix $\mathbf{B}$ are independently generated from standard normal distribution and are then fixed throughout the simulations. In all experiments, we repeat the simulations 500 times. In each run, components of the factor vector $\mathbf{f}_t$ are generated independently from the standard normal distribution. Moreover, we consider three cases of the sparse idiosyncratic component $\mathbf{\Omega}_\varepsilon$, which are similar to those in [8]. More specifically, in Case 1, we consider the chain net-



(a) Student (b) Faculty

Figure 2: Graph indicating the conditional dependence structures of the Webpage example for different categories: left panel, student; right panel, faculty.

work corresponding to the tridiagonal $\mathbf{\Omega}_\varepsilon$ as considered in [5] and [8], which is associated with the auto-regressive covariance structure of order 1. We choose $\mathbf{\Sigma}_u=(\sigma_{ij})$ with $\sigma_{ij}=\exp(-|s_i-s_j|/4)$ in which $s_1<\cdots<s_p$ are generated such that $s_i-s_{i-1}\sim\text{Unif}(0.5,1)$ $(i=2,\cdots,p)$. Then we set $\mathbf{\Omega}_\varepsilon=\mathbf{\Sigma}_\varepsilon^{-1}$. In Case 2, we consider the nearest neighbor networks as in [8]. To generate $\mathbf{\Omega}_\varepsilon$, we first simulate p points from a unit square and calculate all pairwise distances. Then $m=3$ nearest neighbors of each point in $\mathbf{\Omega}_\varepsilon$ are assigned nonzero values, with each of the exact values generated independently from Unif$(0.5,1)$. In Case 3, we generate scale-free networks by using the Barabasi-Albert algorithm [1] for a power-law network, and the nonzero components in the $\mathbf{\Omega}_\varepsilon$ are generated from Unif$(0.5,1)$.

Two loss functions are considered for assessing the performances, the entropy loss (EL) and the Frobenius loss (FL) as follows: $\text{EL}=\text{tr}(\mathbf{\Omega}^{-1}\widehat{\mathbf{\Omega}})-\log\{\det(\mathbf{\Omega}^{-1}\widehat{\mathbf{\Omega}})\}-p$, $\text{FL}=\|\mathbf{\Omega}-\widehat{\mathbf{\Omega}}\|_F^2/\|\mathbf{\Omega}\|_F^2$. Two methods for selecting the tuning parameters are applied, the BIC and cross-validation. The following criteria are calculated for measuring the model selection performances:

$$\text{FP}=\frac{\sum\limits_{1\le j<k\le p} I(\omega_{jk}=0,\hat{\omega}_{jk}\neq 0)}{\sum\limits_{1\le j<k\le p} I(\omega_{jk}=0)},\quad \text{FN}=\frac{\sum\limits_{1\le j<k\le p} I(\omega_{jk}\neq 0,\hat{\omega}_{jk}=0)}{\sum\limits_{1\le j<k\le p} I(\omega_{jk}\neq 0)}.$$

We compare our methods to the naive estimator by simply inverting the sample covariance matrix, and the one obtained by directly applying the CLIME method of [3] without removing the systematic component. In our simulations, the ratio based method is applied to select the number of factors K. In all simulations, the method turns out to work very well with K estimated as 3 consistently.

The results are summarized in Table 1. From Table 1, we can see that using the proposed approach has substantial improvement when compared with other approaches. The BIC and cross-validation has comparable performances under the two losses defined earlier, but BIC performs better in model selection than the cross-validation. As expected, for the two methods that do not exploit the structure in the data model, the one by inverting the sample covariance matrix and the one by directly applying the CLIME method of [3], perform poorly compared with the IPOD approach.

Acknowledgments

Cheng Yong Tang's research was supported by NSF grants IIS-1546087 and SES-1533956. Yingying Fan's research was supported by NSF CAREER Award DMS-1150318. Yinfei Kong's research was supported by NIH R01 DA048176-01 and CSUF Junior/Senior Intramural Grant.

Appendices

A Technical conditions

Let $\mathcal{F}_{-\infty}^{0}$ and $\mathcal{F}_{T}^{\infty}$ be the σ-algebras generated by $\{(\mathbf{y}_t',\mathbf{f}_t')':-\infty\leq t\leq 0\}$ and $\{(\mathbf{y}_t',\mathbf{f}_t')': T\leq t\leq\infty\}$, respectively. Define the mixing coefficient

$$\alpha(T)=\sup_{A\in\mathcal{F}_{-\infty}^{0},B\in\mathcal{F}_{T}^{\infty}}|P(A)P(B)-P(A \text{ and } B)|.$$

Condition 1. The stochastic process $(\mathbf{y}_t',\mathbf{f}_t')'$ is stationary and strong mixing with the mixing coefficient satisfying that for all $T\in\mathbb{Z}^+$,

$$\alpha(T)\leq\exp(-CT^{\gamma_1}),$$

for some constant $\gamma_1>0$.

Let $\mathbf{\Sigma}=\mathbf{U}'\mathbf{\Lambda}\mathbf{U}$ be the eigen-decomposition of the covariance matrix of $\mathbf{y}_t$, where $\mathbf{\Lambda}=\text{diag}\{\lambda_1,\cdots,\lambda_p\}$ is a diagonal matrix with $\lambda_1\geq\lambda_2\geq\cdots\lambda_p\geq 0$ the eigenvalues of $\mathbf{\Sigma}$. Define $\widetilde{\mathbf{Y}}=(\tilde{\mathbf{y}}_1,\cdots,\tilde{\mathbf{y}}_T)=\mathbf{\Lambda}^{-1/2}\mathbf{U}'\mathbf{Y}'$. Then the columns of $\widetilde{\mathbf{Y}}$ have identical distribution $N(\mathbf{0},I_p)$.

Table 1: Simulation results where $\widehat{\Omega}$ is for the proposed approach, $\mathbf{S}^{-1}$ is the inverse of the sample covariance matrix, $\widehat{\Omega}_1$ is the regularized approach of [3] without estimating the systematic component. The sample standard errors are in parentheses.

Case	T			EL	FL	FP $\times 10^2$	FN $\times 10^2$
1	500	$\widehat{\Omega}$	BIC	3.9(0.51)	0.31(0.06)	0.02(0.02)	6.8(0.03)
			CV	3.7(0.43)	0.29(0.04)	43(16)	6.1(0.04)
		$\mathbf{S}^{-1}$		38.1(1.4)	0.87(0.09)	-	-
		$\widehat{\Omega}_1$	BIC	19.6(0.9)	0.63(0.04)	-	-
			CV	17.6(0.8)	0.72(0.06)	-	-
	1000	$\widehat{\Omega}$	BIC	2.6(0.45)	0.24(0.04)	0.01(0.01)	4.6(0.02)
			CV	2.5(0.38)	0.23(0.03)	45(15)	4.2(0.06)
		$\mathbf{S}^{-1}$		21.3(0.8)	0.62(0.05)	-	-
		$\widehat{\Omega}_1$	BIC	18.9(0.6)	0.59(0.03)	-	-
			CV	18.1(0.6)	0.69(0.05)	-	-
2	500	$\widehat{\Omega}$	BIC	3.2(0.46)	0.28(0.04)	0.03(0.02)	6.0(0.02)
			CV	3.1(0.33)	0.25(0.03)	50(18)	6.0(0.04)
		$\mathbf{S}^{-1}$		32.6(1.1)	0.91(0.05)	-	-
		$\widehat{\Omega}_1$	BIC	18.5(0.7)	0.60(0.04)	-	-
			CV	17.5(0.9)	0.71(0.05)	-	-
	1000	$\widehat{\Omega}$	BIC	2.6(0.5)	0.24(0.04)	0.01(0.01)	4.2(0.02)
			CV	2.5(0.5)	0.23(0.03)	45(15)	4.1(0.06)
		$\mathbf{S}^{-1}$		19.3(0.8)	0.87(0.03)	-	-
		$\widehat{\Omega}_1$	BIC	18.2(0.7)	0.57(0.03)	-	-
			CV	17.2(0.7)	0.68(0.04)	-	-
3	500	$\widehat{\Omega}$	BIC	4.5(0.51)	0.35(0.05)	0.02(0.02)	4.3(0.02)
			CV	4.4(0.31)	0.33(0.05)	52(19)	3.9(0.04)
		$\mathbf{S}^{-1}$		32.3(1.2)	0.81(0.06)	-	-
		$\widehat{\Omega}_1$	BIC	18.8(0.9)	0.66(0.04)	-	-
			CV	17.9(0.7)	0.69(0.05)	-	-
	1000	$\widehat{\Omega}$	BIC	2.6(0.53)	0.24(0.04)	0.01(0.01)	3.5(0.02)
			CV	2.7(0.48)	0.23(0.03)	46(17)	3.1(0.06)
		$\mathbf{S}^{-1}$		19.2(0.7)	0.59(0.04)	-	-
		$\widehat{\Omega}_1$	BIC	18.2(0.7)	0.59(0.03)	-	-
			CV	17.2(0.8)	0.64(0.04)	-	-

Condition 2. There exist some positive constants d_j, $j=1,\cdots,K$, and M_1 with $M_1 \geq d_1 \geq \cdots \geq d_K \geq M_1^{-1}$ such that as $p \to \infty$, $\sum_{k=1}^{K} |\lambda_k/p - d_k| = o((\log p)^{-1})$, for any $K < j \leq T$, it holds $M_1^{-1} \leq \lambda_j \leq M_1$, and for $j > p-T$, it holds $0 \leq \lambda_j \leq M_1$.

Condition 2 assumes that the first K eigenvalues of $\mathbf{\Sigma}$ have magnitudes of order p, the next $T-K$ eigenvalues are bounded from both below and above, and the remaining $p-T$ eigenvalues are bounded from above. This assumption is supported by Proposition 1 in [6], which characterizes the orders of λ_j under mild conditions of $\mathbf{B}$ and $\mathbf{\Sigma}_\varepsilon$.

Condition 3. There exists a constant $M_0>0$ such that $|\mathbf{B}|_\infty \le M_0$.

B Supplementary material by C. Y. Tang and Y. Fan

The supplementary material in this appendix contains technical proofs for the main theorems of this paper.

B.1 Lemma B.1 and its proof

Lemma B.1. *Assume conditions of Proposition* 2.1 *hold. If* $\log p = o(T^{\gamma/(2-\gamma)})$ *with* $\gamma=\gamma_1/(1+\gamma_1)$*, then there exist positive constants* c_4 *and* c_5 *such that*

$$P\Big(|\widehat{\mathbf{S}}-I_p|_\infty > c_4\sqrt{T^{-1}\log p}\Big) \le o(p^{-c_5}),$$

where $\widehat{\mathbf{S}}=T^{-1}\widetilde{\mathbf{Y}}'\widetilde{\mathbf{Y}}$ *is the sample covariance matrix.*

Proof. We first prove that for any $i,j\in\{1,2,\cdots,p\}$,

$$P(|\widehat{\mathbf{S}}_{ij}-I_{p,ij}|>c_4\sqrt{T^{-1}\log p}) \le o(p^{-c_5-2}), \tag{B.1}$$

where $\widehat{\mathbf{S}}_{ij}$ and $I_{p,ij}$ are the (i,j) entries of the sample covariance matrix $\widehat{\mathbf{S}}$ and the identity matrix I_p, respectively. Then noting that $P(|\widehat{\mathbf{S}}-I_p|_\infty>x)\le p^2\max_{ij}P(|\widehat{\mathbf{S}}_{ij}-I_{p,ij}|>x)$ for any $x>0$ completes the proof of the theorem. In the following, we use $C_1,C_2,\cdots$ to denote positive generic constants whose values may change from line to line.

We now proceed to prove (B.1). To this end, note that $\widehat{\mathbf{S}}_{ij}=\mathbf{e}'_{p,i}\widehat{\mathbf{S}}\mathbf{e}_{p,j}$, where $\mathbf{e}_{p,j}$ is a p-dimensional unit vector with j-th covariate 1 and all other covariates 0. Let $\tilde{\mathbf{y}}_i=(\tilde{y}_{1i},\cdots,\tilde{y}_{Ti})'=\widetilde{\mathbf{Y}}'\mathbf{e}_{p,i}$. Then $\widehat{\mathbf{S}}_{ij}=T^{-1}\tilde{\mathbf{y}}_i'\tilde{\mathbf{y}}_j$ and $\tilde{\mathbf{y}}_i$ is a T-dimensional random vector whose elements $\tilde{y}_{ti}$ have standard norm distribution. Moreover, for each $t=1,\cdots,T$, the random variables $\tilde{y}_{ti}$ and $\tilde{y}_{tj}$, $i\neq j$ are independent. Then by Gaussian tail probability, we know that for any $u>0$,

$$P(|\tilde{y}_{ti}\tilde{y}_{sj}|>u)\le P(|\tilde{y}_{ti}|>\sqrt{u})+P(|\tilde{y}_{sj}|>\sqrt{u})\le C_1\exp(-C_2u).$$

By Condition 1, applying the Bernstein's inequality for weakly dependent sequence (Theorem 1 of [14]) yields

$$\begin{aligned}&P\Big(|T^{-1}\tilde{\mathbf{y}}_i'\tilde{\mathbf{y}}_j-I_{p,ij}|\geq u\Big)\\ \leq& T\exp\Big(-\frac{(Tu)^\gamma}{C_1}\Big)+\exp\Big(-\frac{T^2u^2}{C_2(1+TC_3)}\Big)+\exp\Big(-\frac{(Tu)^2}{C_4T}\exp\Big(\frac{(Tu)^{\gamma(1-\gamma)}}{C_5(\log(Tu))^\gamma}\Big)\Big),\end{aligned}$$

where $\gamma=\gamma_1/(1+\gamma_1)$. Let $u=c_4\sqrt{(\log p)/T}$ with c_4 some large positive constant, then it follows from the above inequality that if $\log p=o(T^{\gamma/(2-\gamma)})$,

$$P(|T^{-1}\tilde{\mathbf{y}}_i'\tilde{\mathbf{y}}_j-I_{p,ij}|\geq c_4\sqrt{(\log p)/T})=o(p^{-c_5-2}).$$

This proves (B.1) and thus completes the proof of the lemma. □

B.2 Lemma B.2 and its proof

Lemma B.2. *Assume Conditions* 2-3 *hold. Then we have*

$$|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}|_\infty=O_p\Big(K^2\sqrt{\frac{\log K}{T}}+K\sqrt{\frac{\log p}{T}}\Big).$$

Proof. Using model (2.2), we have the decomposition

$$\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}=\mathbf{D}_1+\mathbf{D}_2+\mathbf{D}_3+\mathbf{D}_3', \tag{B.2}$$

where $\mathbf{D}_1=T^{-1}\mathbf{B}\big(\sum_{t=1}^T\mathbf{f}_t\mathbf{f}_t'-I_K\big)\mathbf{B}'$, $\mathbf{D}_2=T^{-1}\sum_{t=1}^T(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'-\mathbf{\Sigma}_\varepsilon)$, and $\mathbf{D}_3=T^{-1}\mathbf{B}\sum_{t=1}^T\mathbf{f}_t\boldsymbol{\varepsilon}_t'$. In the following, we use $C_1,C_2,\cdots$ to denote generic positive constants whose values may change from line to line.

We first consider $\mathbf{D}_1$. Let $\mathbf{e}_{p,j}$ be a p-dimensional unit vector with j-th covariate 1 and all other covariates 0. Since $\mathbf{f}_t\sim N(\mathbf{0},I_K)$ and $T^{-1}\sum_{t=1}^T\mathbf{f}_t\mathbf{f}_t'$ is the sample covariance matrix estimate of I_K, by Condition 1 and using similar proof as for Lemma B.1 we obtain that with probability at least $1-o(T^{-C_1})$,

$$\Big|T^{-1}\sum_{t=1}^T\mathbf{f}_t\mathbf{f}_t'-I_K\Big|_\infty\leq C_2\sqrt{(\log T)/T}.$$

Thus by the Cauchy-Schwarz inequality and Condition 3, we have

$$\begin{aligned}|\mathbf{D}_1|_\infty=&\max_{1\leq i,j\leq p}|\mathbf{e}_{p,i}'\mathbf{D}_1\mathbf{e}_{p,j}|\leq\Big\|T^{-1}\sum_{t=1}^T\mathbf{f}_t\mathbf{f}_t'-I_K\Big\|_2\max_{1\leq i,j\leq p}\|\mathbf{B}'\mathbf{e}_{p,i}\|_2\|\mathbf{B}'\mathbf{e}_{p,j}\|_2\\ \leq&\mathcal{O}(K)\Big\|T^{-1}\sum_{t=1}^T\mathbf{f}_t\mathbf{f}_t'-I_K\Big\|_2=O_p\Big(K^2\sqrt{(\log T)/T}\Big),\end{aligned} \tag{B.3}$$

where the last step is because of

$$\Big\|T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t'-I_K\Big\|_2\le K\Big|T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t'-I_K\Big|_\infty=O_p\big(K\sqrt{(\log T)/T}\big).$$

Next we consider $\mathbf{D}_2$. Since $\boldsymbol{\varepsilon}_t=\mathbf{y}_t-\mathbf{B}\mathbf{f}_t$, by Condition 1 and using similar proof as that for Lemma B.1 we obtain that with probability at least $1-o(p^{-C_3})$,

$$|\mathbf{D}_2|_\infty\le C_4\sqrt{(\log p)/T}. \tag{B.4}$$

Finally, we study $\mathbf{D}_3$. First, following the similar proof as that for Lemma B.1 we obtain that

$$\max_{1\le i\le K,1\le j\le p}\Big|T^{-1}\sum_{t=1}^{T}f_{it}u_{jt}\Big|=O_p\big(\sqrt{(\log p)/T}\big).$$

Thus, by the Cauchy-Schwarz inequality and Condition 3, we have

$$\begin{aligned}|\mathbf{D}_3|_\infty&\le\max_{1\le i,j\le p}\|\mathbf{B}'\mathbf{e}_{p,i}\|_2\Big\|T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\boldsymbol{\varepsilon}_t'\mathbf{e}_{p,j}\Big\|_2\le\sqrt{K}\max_{1\le j\le p}\Big\|T^{-1}\sum_{t=1}^{T}\mathbf{f}_t\boldsymbol{\varepsilon}_t'\mathbf{e}_{p,j}\Big\|_2\\&\le K\max_{1\le l,j\le p}\Big|T^{-1}\mathbf{e}_{K,l}'\sum_{t=1}^{T}\mathbf{f}_t\boldsymbol{\varepsilon}_t'\mathbf{e}_{p,j}\Big|\le O_p(K\sqrt{(\log p)/T}).\end{aligned} \tag{B.5}$$

Combining (B.2)-(B.5) and comparing and collecting the terms complete the proof of the lemma. □

B.3 Proof of Proposition 2.1

The key idea is to consider the dual matrix $\mathbf{S}_D=T^{-1}\mathbf{Y}\mathbf{Y}'$, whose eigenvalues are the same as the nonzero eigenvalues of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}=T^{-1}\mathbf{Y}'\mathbf{Y}$. Recall that $\widetilde{\mathbf{Y}}=(\tilde{y}_{jt})=\boldsymbol{\Lambda}^{-1/2}\mathbf{U}'\mathbf{Y}'$, whose columns have identical distribution $N(\mathbf{0},I_p)$. Then the dual matrix can be written as $\mathbf{S}_D=T^{-1}\widetilde{\mathbf{Y}}'\boldsymbol{\Lambda}\widetilde{\mathbf{Y}}$. In the proof below, we use $C_1,C_2,\cdots$ to denote some generic positive constants. For each i, we define $\phi_i(\cdot)$ as the function which takes out the ith largest eigenvalue of a matrix.

Let $\mathbf{D}$ be a $p\times p$ diagonal matrix with the first K diagonal elements being equal to d_j, and the rest being 1. We decompose $\widetilde{\mathbf{Y}}$, $\mathbf{D}$ and $\boldsymbol{\Lambda}$ as

$$\widetilde{\mathbf{Y}}=(\widetilde{\mathbf{Y}}_1',\widetilde{\mathbf{Y}}_2')',\quad\mathbf{D}=\mathrm{diag}\{\mathbf{D}_1,I_{p-K}\},\quad\boldsymbol{\Lambda}=\mathrm{diag}\{\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2\},$$

where $\widetilde{\mathbf{Y}}_1,\mathbf{D}_1$ and $\mathbf{\Lambda}_1$ all have K rows. Then it follows that

$$p^{-1}\mathbf{S}_D=(Tp)^{-1}\widetilde{\mathbf{Y}}'\mathbf{\Lambda}\widetilde{\mathbf{Y}}=(Tp)^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1+(Tp)^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2. \tag{B.6}$$

By Weyl's inequality [9] we have for any $j=1,\cdots,T$,

$$\begin{aligned}&\phi_j(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1)+\phi_T(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2)\le\phi_j(p^{-1}\mathbf{S}_D)\\ \le&\phi_j(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1)+\phi_1(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2).\end{aligned}$$

The above inequality together with $\phi_T(\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2)\ge 0$ entails that for $j=1,\cdots,T$,

$$\begin{aligned}\phi_j(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1)&\le\phi_j(p^{-1}\mathbf{S}_D)\\ &\le\phi_1(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2)+\phi_j(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1).\end{aligned} \tag{B.7}$$

Applying Weyl's inequality [9] one more time we obtain for any $j=K+1,\cdots,T$,

$$\begin{aligned}\phi_j(p^{-1}\mathbf{S}_D)&\ge\phi_T(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1)+\phi_j(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2)\\ &=\phi_j(p^{-1}T^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2),\end{aligned} \tag{B.8}$$

where the last step is because $\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1$ is positive semidefinite and has rank K, which is much smaller than T.

We will prove that with probability at least $1-\mathcal{O}(T^{-C_3})$, for $j=1,\cdots,K$,

$$\frac{2}{3}M_1^{-1}<\phi_j(T^{-1}p^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1)<\frac{7}{3}M_1. \tag{B.9}$$

And with probability at least $1-\mathcal{O}(p^{-C_4})$,

$$\phi_1(T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2)\le 2M_1p^{-1}+C_5M_1\sqrt{T^{-1}\log p}. \tag{B.10}$$

Since $\phi_j(\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1)=0$ for $j=K+1,\cdots,T$ and $\hat{\lambda}_j=\phi_j(\widehat{\mathbf{\Sigma}})=\phi_j(\mathbf{S}_D)$ for $j=1,\cdots,T$, combining (B.7)-(B.10) proves that with probability at least $1-\mathcal{O}(T^{-C_3})$,

$$\frac{2}{3}M_1^{-1}\le p^{-1}\hat{\lambda}_j\le\frac{7}{3}M_1,\quad\text{for } j=1,\cdots,K, \tag{B.11}$$

$$p^{-1}\hat{\lambda}_j\le 2M_1p^{-1}+C_5M_1\sqrt{T^{-1}\log p},\quad\text{for } j=K+1,\cdots,T. \tag{B.12}$$

This completes the proof of Proposition 2.1.

We now proceed to prove (B.9). Define event $\mathcal{E}=\cap_{t=1}^{T}\cap_{j=1}^{p}\{|\tilde{y}_{jt}|\le C_1\sqrt{\log p}\}$ with $C_1>0$ some large enough positive constant. Since $\tilde{y}_{jt}\sim N(0,1)$, it follows from Gaussian tail probability that

$$P(\mathcal{E}^c)\le\sum_{t=1}^{T}\sum_{j=1}^{p}P(|\tilde{y}_{jt}|>C_1\sqrt{\log p})\le o(p^{-C_2}).$$

This together with the assumption $\sum_{k=1}^{K}|\lambda_k/p-d_k|=o((\log p)^{-1})$ ensures that with probability at least $1-o(p^{-C_2})$,

$$\begin{aligned}|\widetilde{\mathbf{Y}}_1'(p^{-1}\mathbf{\Lambda}_1-\mathbf{D}_1)\widetilde{\mathbf{Y}}_1|_\infty&=\max_{1\le j,\ell\le T}\Big|\sum_{k=1}^{K}(\lambda_k/p-d_k)\tilde{y}_{jk}\tilde{y}_{k\ell}\Big|\\&\le C_1^2(\log p)\sum_{k=1}^{K}|\lambda_k/p-d_k|\to 0.\end{aligned}$$

Since for any $T\times T$ matrix $\mathbf{A}$, we have $\|\mathbf{A}\|_F\le T|\mathbf{A}|_\infty$, it follows that

$$\|T^{-1}\widetilde{\mathbf{Y}}_1'(p^{-1}\mathbf{\Lambda}_1-\mathbf{D}_1)\widetilde{\mathbf{Y}}_1\|_F\le|\widetilde{\mathbf{Y}}_1'(p^{-1}\mathbf{\Lambda}_1-\mathbf{D}_1)\widetilde{\mathbf{Y}}_1|_\infty\to 0. \tag{B.13}$$

Further, by Corollary 6.3.8 of [9] we obtain

$$\max_{1\le i\le K}\Big|\phi_i\big((Tp)^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{\Lambda}_1\widetilde{\mathbf{Y}}_1\big)-\phi_i\big(T^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{D}_1\widetilde{\mathbf{Y}}_1\big)\Big|\le\|T^{-1}\widetilde{\mathbf{Y}}_1'(p^{-1}\mathbf{\Lambda}_1-\mathbf{D}_1)\widetilde{\mathbf{Y}}_1\|_F\to 0. \tag{B.14}$$

Meanwhile, for $1\le i\le K$, similar to (B.7) we can prove

$$\begin{aligned}\phi_T\big(T^{-1}\widetilde{\mathbf{Y}}_1'(\mathbf{D}_1-d_KI_K)\widetilde{\mathbf{Y}}_1\big)&\le\phi_i\big(T^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{D}_1\widetilde{\mathbf{Y}}_1\big)-d_K\phi_i\big(T^{-1}\widetilde{\mathbf{Y}}_1'\widetilde{\mathbf{Y}}_1\big)\\&\le\phi_1\big(T^{-1}\widetilde{\mathbf{Y}}_1'(\mathbf{D}_1-d_KI_K)\widetilde{\mathbf{Y}}_1\big).\end{aligned}$$

Since the matrix $T^{-1}\widetilde{\mathbf{Y}}_1'(\mathbf{D}_1-d_KI_K)\widetilde{\mathbf{Y}}_1$ is positive semidefinite and has rank K, it follows that $\phi_T\big(T^{-1}\widetilde{\mathbf{Y}}_1'(\mathbf{D}_1-d_KI_K)\widetilde{\mathbf{Y}}_1\big)=0$. Meanwhile, since $\phi_1\big(T^{-1}\widetilde{\mathbf{Y}}_1'(\mathbf{D}_1-d_KI_k)\widetilde{\mathbf{Y}}_1\big)\le\max_{1\le j\le K}|d_j-d_K|\phi_1\big(T^{-1}\widetilde{\mathbf{Y}}_1'\widetilde{\mathbf{Y}}_1\big)\le M_1\phi_1\big(T^{-1}\widetilde{\mathbf{Y}}_1'\widetilde{\mathbf{Y}}_1\big)$ and $M_1^{-1}\le d_i\le M_1$, we have

$$M_1^{-1}\phi_i\big(T^{-1}\widetilde{\mathbf{Y}}_1'\widetilde{\mathbf{Y}}_1\big)\le\phi_i\big(T^{-1}\widetilde{\mathbf{Y}}_1'\mathbf{D}_1\widetilde{\mathbf{Y}}_1\big)\le 2M_1\phi_1\big(T^{-1}\widetilde{\mathbf{Y}}_1'\widetilde{\mathbf{Y}}_1\big). \tag{B.15}$$

So the key is to study the first K eigenvalues of matrix $T^{-1}\widetilde{\mathbf{Y}}_1'\widetilde{\mathbf{Y}}_1$, which are the same as eigenvalues of its $K\times K$ dual matrix $T^{-1}\widetilde{\mathbf{Y}}_1\widetilde{\mathbf{Y}}_1'$. Notice that the columns of $\widetilde{\mathbf{Y}}_1$ have identical distribution $N(\mathbf{0},I_K)$. Then using similar proof to that for Lemma B.1, we can prove that with probability at least $1-\mathcal{O}(T^{-C_3})$,

$$|T^{-1}\widetilde{\mathbf{Y}}_1\widetilde{\mathbf{Y}}_1'-I_K|_\infty\leq C_4\sqrt{T^{-1}\log T}.$$

This ensures that if $K=o(\sqrt{T/\log T})$, by Corollary 6.3.8 of [9], we have with probability at least $1-\mathcal{O}(T^{-C_3})$,

$$\begin{aligned}&\max_{1\leq i\leq T}|\phi_i(T^{-1}\widetilde{\mathbf{Y}}_1\widetilde{\mathbf{Y}}_1')-1|\leq\|T^{-1}\widetilde{\mathbf{Y}}_1\widetilde{\mathbf{Y}}_1'-I_K\|_F\\ \leq&K|T^{-1}\widetilde{\mathbf{Y}}_1\widetilde{\mathbf{Y}}_1'-I_K|_\infty\leq C_4K\sqrt{T^{-1}\log T}\rightarrow 0.\end{aligned}$$

Combing the above results with (B.14) and (B.15) and noting $\phi_j(T^{-1}\widetilde{\mathbf{Y}}_1\widetilde{\mathbf{Y}}_1')=\phi_j(T^{-1}\widetilde{\mathbf{Y}}_1'\widetilde{\mathbf{Y}}_1)$ for $j=1,\cdots,K$ completes the proof of (B.9).

Next we prove (B.10). Similar to (B.15) we can prove for $j=1,\cdots,T$,

$$M_1^{-1}\phi_T(T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2)\leq\phi_j(T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2)\leq 2M_1\phi_1(T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2).\tag{B.16}$$

So the key is to study the eigenvalues of matrix $T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2$. Similar to the previous $j=1,\cdots,K$ case, we only need to study the eigenvalues of $T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2\widetilde{\mathbf{Y}}_2'$. Notice that the columns of $\widetilde{\mathbf{Y}}_2$ have identical distribution $N(\mathbf{0},I_{p-K})$. Thus, by Lemma B.1, we obtain that with probability at least $1-\mathcal{O}(p^{-c_4})$,

$$|T^{-1}\widetilde{\mathbf{Y}}_2\widetilde{\mathbf{Y}}_2'-I_{p-K}|_\infty\leq c_3\sqrt{T^{-1}\log p},\tag{B.17}$$

where c_3 and c_4 are defined in Lemma B.1. Therefore, we can derive that if $\sqrt{T^{-1}\log p}\rightarrow 0$,

$$\begin{aligned}\|T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2\widetilde{\mathbf{Y}}_2'\|_F&\leq p^{-1}\|T^{-1}\widetilde{\mathbf{Y}}_2\widetilde{\mathbf{Y}}_2'-I_{p-K}\|_F+p^{-1}\|I_{p-K}\|_F\\ &\leq c_3\sqrt{T^{-1}\log p}+p^{-1/2}\rightarrow 0.\end{aligned}$$

By Corollary 6.3.8 of [9], we have

$$\phi_1(T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2)\leq\|T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2\|_F\leq c_3\sqrt{T^{-1}\log p}+p^{-1/2}\rightarrow 0.$$

This together with (B.16) entails that (B.10) holds with $C_4=c_4$ and $C_5=c_3$. This completes the proof of the proposition. □

B.4 Proof of Theorem 2.1

Proof. We only need to prove that with probability at least $1-\mathcal{O}(p^{-C_6})$, for $j=1,\cdots,T$,

$$1/2\le\phi_j(p^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2)\le 2. \tag{B.18}$$

Then this together with (B.16) ensures that for $j=1,\cdots,T$,

$$(2M_1T)^{-1}\le\phi_j(T^{-1}p^{-1}\widetilde{\mathbf{Y}}_2'\mathbf{\Lambda}_2\widetilde{\mathbf{Y}}_2)\le 4M_1T^{-1}.$$

In view of (B.7) and (B.8), for $j=K+1,\cdots,T$,

$$(2M_1T)^{-1}\le\hat{\lambda}_j=\phi_j(p^{-1}\mathbf{S}_D)\le 4M_1T^{-1}. \tag{B.19}$$

Note that $\hat{\lambda}_j=\phi_j(\widehat{\mathbf{\Sigma}})=\phi_j(\mathbf{S}_D)$ for $j=1,\cdots,T$. Combining (B.19) with (B.11) proves that in addition to (2.13), it also holds that for $j=K+1,\cdots,T$,

$$\hat{\lambda}_j/\hat{\lambda}_{j+1}<\tilde{c}_9,$$

and for $j=K$,

$$\hat{\lambda}_K/\hat{\lambda}_{K+1}\ge c_{10}T\to\infty,$$

where $\tilde{c}_9$ and c_{10} are two positive constants. This completes the proof of the theorem.

It remains to prove (B.18). In the following, we use $\tilde{C}_1,\tilde{C}_2$ to denote some positive generic constants. Since $\{\mathbf{y}_t\}_{1\le t\le T}$ are independent across t, the rows of $\widetilde{\mathbf{Y}}_2$ are independent. Recall that the columns of $\widetilde{\mathbf{Y}}_2$ have identical distribution $N(\mathbf{0},I_{p-K})$. Thus, the entries of matrix $\widetilde{\mathbf{Y}}_2$ are independent standard normal random variables. This ensures that $(p-K)^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2$ is the sample estimate of the covariance matrix I_T. By Lemma B.1 we have with probability at least $1-\mathcal{O}(T^{-\tilde{C}_1})$,

$$|(p-K)^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2-I_T|_\infty\le\tilde{C}_2\sqrt{p^{-1}\log T}.$$

Thus, by Corollary 6.3.8 of [9], we have

$$\max_{1\le j\le T}|\phi_j((p-K)^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2)-1|\le\|\phi_j((p-K)^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2)-1\|_F\le\tilde{C}_2T\sqrt{p^{-1}\log T}\to 0.$$

It follows from the above inequality that for T large enough, with probability at least $1-\mathcal{O}(T^{-\tilde{C}_1})$,

$$1/2\le\phi_j((p-K)^{-1}\widetilde{\mathbf{Y}}_2'\widetilde{\mathbf{Y}}_2)\le 2.$$

This together with $p/(p-K)\to 1$ completes the proof of (B.18). This concludes the proof of Theorem 2.1. □

B.5 Proof of Lemma 2.1

By (2.5), (2.10) and the Cauchy-Schwarz inequality, we have

$$|\widehat{\mathbf{\Sigma}}_\varepsilon-\mathbf{\Sigma}_\varepsilon|_\infty\le|\widehat{\mathbf{B}}\widehat{\mathbf{B}}'-\mathbf{B}\mathbf{B}'|_\infty+|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}|_\infty.$$

The first term $\widehat{\mathbf{B}}\widehat{\mathbf{B}}'$ on the right-hand side has been studied in [6] in the proof of Theorem 3.2 (page 43 therein). It was proved that

$$|\widehat{\mathbf{B}}\widehat{\mathbf{B}}'-\mathbf{B}\mathbf{B}'|_\infty=O_p\Big(K^3\sqrt{(\log K)/T}+\delta_T^*\Big), \tag{B.20}$$

where

$$\delta_T^*=\frac{K\sqrt{\log p}+K^2}{\sqrt{T}}+\frac{K^3}{\sqrt{p}}+\sqrt{\frac{\log p}{T}}.$$

The second term has been studied in Lemma B.2. Thus, combining these two results completes the proof of the lemma.

B.6 Proof of Theorem 2.2

The results follow directly from Theorem 6 of [3].

B.7 Proof of Theorem 2.3

The proof is similar to the one for Theorem 3.2 in [6]. Note that

$$\widehat{\mathbf{\Omega}}-\mathbf{\Omega}=L_1+L_2'+L_3+L_4'+L_5+L_6,$$

where the L_i's are defined as follows:

$$\begin{aligned}
&L_1=\widehat{\mathbf{\Omega}}_\varepsilon-\mathbf{\Omega}_\varepsilon,\\
&L_2=(\widehat{\mathbf{\Omega}}_\varepsilon-\mathbf{\Omega}_\varepsilon)\widehat{\mathbf{B}}[I_K+\widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon,\\
&L_3=(\widehat{\mathbf{\Omega}}_\varepsilon-\mathbf{\Omega}_\varepsilon)\widehat{\mathbf{B}}[I_K+\widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\mathbf{\Omega}_\varepsilon,\\
&L_4=\mathbf{\Omega}_\varepsilon(\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H}^{-1})[I_K+\widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\mathbf{\Omega}_\varepsilon,\\
&L_5=\mathbf{\Omega}_\varepsilon(\widehat{\mathbf{B}}-\mathbf{B}\mathbf{H}^{-1})[I_K+\widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}(\mathbf{H}')^{-1}\mathbf{B}'\mathbf{\Omega}_\varepsilon,\\
&L_6=\mathbf{\Omega}_\varepsilon\mathbf{B}\mathbf{H}^{-1}\big([I_K+\widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}-[\mathbf{H}'\mathbf{H}+(\mathbf{H}')^{-1}\mathbf{B}'\mathbf{\Omega}_\varepsilon\mathbf{B}\mathbf{H}^{-1}]^{-1}\big)(\mathbf{H}')^{-1}\mathbf{B}'\mathbf{\Omega}_\varepsilon,
\end{aligned}$$

with $\mathbf{H}=\frac{1}{T}\mathbf{U}^{-1}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\mathbf{B}$ and $\mathbf{U}=\mathrm{diag}\{\widehat{\lambda}_1,\cdots,\widehat{\lambda}_K\}$.

Let $\tilde{\delta}_T = s_p \delta_T^{1-q}$. Then, by Theorem 2.2, we have $\|\widehat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}_\varepsilon\|_2 = O_p(\tilde{\delta}_T)$. Using the same proof as the one for Theorem 3.2 in [6], we have

$$\|L_1\|_2 = \|\widehat{\mathbf{\Omega}}_\varepsilon - \mathbf{\Omega}_\varepsilon\|_2 = O_p(\tilde{\delta}_T), \quad \|L_2\|_2 = O_p(\|L_1\|_2), \quad \|L_3\|_2 = O_p(\|L_1\|_2),$$
$$\|L_4\|_2 = O_p\Big(\sqrt{K^5 p^{-1}} + \sqrt{\frac{\log p + K^3}{T}}\Big), \quad \|L_5\|_2 = O_p(\|L_4\|_2), \quad \|L_6\|_2 = O_p(\tilde{\delta}_T).$$

Combining the above results and comparing the orders of terms we have

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 \le O_p\big(\tilde{\delta}_T + \sqrt{K^5 p^{-1}} + \sqrt{(\log p + K^3)/T}\big) = O_p(\tilde{\delta}_T), \tag{B.21}$$

which completes the proof of the first result.

Finally we study the estimation error under the Frobenius norm. We will repeatedly use the following inequality stated in exercise 20 on page 313 of [9]:

$$\|\mathbf{A}_1\mathbf{A}_2\|_F \le \|\mathbf{A}_1\|_2 \|\mathbf{A}_2\|_F \quad \text{and} \quad \|\mathbf{A}_1\mathbf{A}_2\|_F \le \|\mathbf{A}_1\|_F \|\mathbf{A}_2\|_2, \tag{B.22}$$

for all matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ of the proper sizes. In addition, it has been shown in [6] that

$$\begin{aligned}
&\|(I_K + \widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}})^{-1}\|_2 = O_p(p^{-1}), \quad \|\widehat{\mathbf{B}}\|_2 = O_p(\sqrt{p}),\\
&\|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}^{-1}\|_F^2 = O_p(K^5 + T^{-1} p \log p + T^{-1} p K^3),\\
&\max\{\|\mathbf{H}\|_2, \|\mathbf{H}^{-1}\|_2\} = O_p(1).
\end{aligned}$$

By Theorem 2.2, $\|L_1\|_F^2 \le p s_0(p) O_p(\delta_T^{2-q})$. Similarly to the proof of Theorem 3.2 (p.40) in [6], we can prove that

$$\begin{aligned}
\|L_2\|_F^2 &\le \|L_1\|_F^2 \|\widehat{\mathbf{B}}[I_K + \widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\|_2^2 = O_p(\|L_1\|_F^2), && \text{(B.23)}\\
\|L_3\|_F^2 &\le \|L_1\|_F^2 \|\widehat{\mathbf{B}}[I_K + \widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\mathbf{\Omega}_\varepsilon\|_2^2 = O_p(\|L_1\|_F^2), && \text{(B.24)}\\
\|L_4\|_F^2 &\le \|\mathbf{\Omega}_\varepsilon\|_2^2 \|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}^{-1}\|_F^2 \|[I_K + \widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\widehat{\mathbf{B}}'\mathbf{\Omega}_\varepsilon\|_2^2 &&\\
&\le \|\mathbf{\Omega}_\varepsilon\|_2^2 \|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}^{-1}\|_F^2 \|[I_K + \widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\|_2^2 \|\widehat{\mathbf{B}}'\|_2^2 \|\mathbf{\Omega}_\varepsilon\|_2^2 &&\\
&\le O_p(K^5/p + (\log p + K^3)/T), && \text{(B.25)}\\
\|L_5\|_F^2 &\le \|\mathbf{\Omega}_\varepsilon\|_2^2 \|\widehat{\mathbf{B}} - \mathbf{B}\mathbf{H}^{-1}\|_F^2 \|[I_K + \widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}\|_2^2 \|(\mathbf{H}')^{-1}\mathbf{B}'\mathbf{\Omega}_\varepsilon\|_2^2, &&\\
&\le O_p(K^5/p + (\log p + K^3)/T). && \text{(B.26)}
\end{aligned}$$

Finally we study $\|L_6\|_F^2$. Let $\mathbf{G} = [I_K + \widehat{\mathbf{B}}'\widehat{\mathbf{\Omega}}_\varepsilon\widehat{\mathbf{B}}]^{-1}$ and $\mathbf{G}_1 = [\mathbf{H}'\mathbf{H} + (\mathbf{H}')^{-1}\mathbf{B}'\mathbf{\Omega}_\varepsilon\mathbf{B}\mathbf{H}^{-1}]^{-1}$. Then similar to the proof of Theorem 3.2 in [6] (p.40)

$\|\mathbf{G}\|_2=O_p(p^{-1})=\|\mathbf{G}_1\|_2$, and

$$\|\mathbf{G}-\mathbf{G}_1\|_F\le\sqrt{K}\|\mathbf{G}-\mathbf{G}_1\|_2\le O_p(\sqrt{K}p^{-1}\tilde{\delta}_T).$$

Therefore,

$$\|L_6\|_F^2\le\|\mathbf{\Omega}_\varepsilon\mathbf{B}\mathbf{H}^{-1}\|_2^4\|\mathbf{G}-\mathbf{G}_1\|_F^2=O_p(K\tilde{\delta}_T^2). \tag{B.27}$$

Combining (B.23)-(B.27) we obtain

$$\begin{aligned}p^{-1}\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}\|_F^2&\le O_p\Big(s_p\delta_T^{2-q}+\frac{K^5}{p^2}+\frac{\log p+K^3}{pT}+\frac{K\tilde{\delta}_T^2}{p}\Big)\\&=O_p\Big(s_p\delta_T^{2-q}+\frac{Ks_p^2\delta_T^{2-2q}}{p}+\frac{K^5}{p^2}+\frac{\log p+K^3}{pT}\Big),\end{aligned}$$

which concludes the proof of the theorem.

References

[1] Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

[2] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics*, 36:199–227.

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

[4] Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40:1935–1967.

[5] Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *Annals of Applied Statistics*, 3:521–541.

[6] Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B*, 75:603–680.

[7] Friedman, J., Hastie, T., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–411.

[8] Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98:1–15.

[9] Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. New York: Cambridge University Press.

[10] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37:4254–4278.

[11] Lam, C. and Yao, Q. (2012). Factor modelling for high-dimensional time series: inference for the number of factors. *Annals of Statistics*, 40:694–726.

[12] Lauritzen, S. L. (1996). *Graphical Models*. New York: Oxford University Press.

[13] Leng, C. and Tang, C. Y. (2012). Sparse matrix graphical models. *Journal of the American Statistical Association*, 107:1187–1200.

[14] Merlevéde, F., Peligrad, M. and Rio, E. (2009). A Bernstein type inequality and moderate deviations for weakly dependent sequences. Manuscript, Université Paris Est.

[15] Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104:735–746.

[16] Rothman, A., Levina, L., and Zhu, J. (2010). A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika*, 97:539–550.

[17] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

[18] Seung, C. A. and Horenstein, A.R. (2013). Eigenvalue Ratio Test for the Number of Factors. *Econometrica*, 81: 1203–1227.

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