Embracing the Blessing of Dimensionality in Factor Models

Quefeng Li, Guang Cheng, Jianqing Fan and Yuyan Wang *

October 18, 2016

Abstract

Factor modeling is an essential tool for exploring intrinsic dependence structures among high-dimensional random variables. Much progress has been made for estimating the covariance matrix from a high-dimensional factor model. However, the blessing of dimensionality has not yet been fully embraced in the literature: much of the available data is often ignored in constructing covariance matrix estimates. If our goal is to accurately estimate a covariance matrix of a set of targeted variables, shall we employ additional data, which are beyond the variables of interest, in the estimation? In this paper, we provide sufficient conditions for an affirmative answer, and further quantify its gain in terms of Fisher information and convergence rate. In fact, even an oracle-like result (as if all the factors were known) can be achieved when a sufficiently large number of variables is used. The idea of utilizing data as much as possible brings computational challenges. A divide-and-conquer algorithm is thus proposed to alleviate the computational burden, and also shown not to sacrifice any statistical accuracy in comparison with a pooled analysis. Simulation studies further confirm our advocacy for the use of full data, and demonstrate the effectiveness of the above algorithm. Our proposal is applied to a microarray data example that shows empirical benefits of using more data.

Keywords: Asymptotic normality, auxiliary data, divide-and-conquer, factor model, Fisher information, high-dimensionality.

*Quefeng Li is Assistant Professor, Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599 (Email:quefeng@email.unc.edu). Guang Cheng is Associate Professor, Department of Statistics, Purdue University, IN 47096 (Email:chengg@stat.purdue.edu). Jianqing Fan is Frederick L. Moore '18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, NJ 08544 (Email:jqfan@princeton.edu) and Honorary Professor, School of Data Science, Fudan University, Shanghai, China. Yuyan Wang is PhD student, Department of Operations Research and Financial Engineering, Princeton University, NJ 08544 (Email: yuyan@princeton.edu). Guang Cheng gratefully acknowledges NSF CAREER Award DMS-1151692, DMS-1418042, Simons Fellowship in Mathematics, Office of Naval Research (ONR N00014-15-1-2331). Jianqing Fan was supported in part by NSF Grants DMS-1206464, DMS-1406266, and NIH grant R01GM100474-4. Quefeng Li was supported by NIH grant 2R01-GM072611-11 as a postdoctoral fellow at Princeton University and supported by NIH grant 2R01-GM047845-25 at UNC-Chapel Hill. Guang Cheng was on sabbatical at Princeton while this work was carried out and thanks the Princeton ORFE department for its hospitality. The authors thank the editor, the associate editor, and the referees for their constructive comments.

1 Introduction

With the advance of modern information technology, it is now possible to track millions of variables or subjects simultaneously. To discover the relationship among them, the estimation of a high-dimensional covariance matrix $\mathbf{\Sigma}$ has recently received a great deal of attention in the literature. Researchers proposed various regularization methods to obtain consistent estimators of $\mathbf{\Sigma}$ (Bickel and Levina, 2008; Lam and Fan, 2009; Rothman et al., 2008; Cai and Liu, 2011; Cai et al., 2010). A key assumption for these regularization methods is that $\mathbf{\Sigma}$ is sparse, i.e. many elements of $\mathbf{\Sigma}$ are small or exactly zero.

Different from such a sparsity condition, factor analysis assumes that the intrinsic dependence is mainly driven by some common latent factors (Johnson and Wichern, 1992). For example, in modeling stock returns, Fama and French (1993) proposed the well-known Fama-French three-factor model. In the factor model, $\mathbf{\Sigma}$ has spiked eigenvalues and dense entries. In the high dimensional setting, there are many recent studies on the estimation of the covariance matrix based on the factor model (Fan et al., 2008, 2011, 2013; Bai and Li, 2012; Bai and Liao, 2013), where the number of variables can be much larger than the number of observations.

The interest of this paper is on the estimation of the covariance matrix for a certain set of variables using auxiliary data information. In the literature, we use only the data information on the variables of interest. In the data-rich environment today, substantially more amount of data information is indeed available, but is often ignored in statistical analysis. For example, we might be interested in understanding the covariance matrix of 50 stocks in a portfolio, yet the available data information are a time series of thousands of stocks. Similarly, an oncologist may wish to study the dependence or network structures among 100 genes that are significantly associated with a certain cancer, yet she has expression data for over 20,000 genes from the whole genome. Can we benefit from using much more rich auxiliary data?

The answer to the above question is affirmative when a factor model is imposed. Since the whole system is driven by a few common factors, these common factors can be inferred more accurately from a much larger set of data information (Fan et al., 2013), which is indeed a "blessing of dimensionality". A major contribution of this paper is to characterize how much the estimation of the covariance matrix of interest and also common factors can be improved by auxiliary data information (and under what conditions).

Consider the following factor model for all p observable data $\mathbf{y}_t = (y_{1t}, \ldots, y_{pt})' \in \mathbb{R}^p$ at time t:

$$\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t, \quad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{f}_t \in \mathbb{R}^K$ is a K-dimensional vector of common factors, $\mathbf{B} = (\mathbf{b}_1', \ldots, \mathbf{b}_p')' \in \mathbb{R}^{p\times K}$ is a factor loading matrix with $\mathbf{b}_i \in \mathbb{R}^K$ being the factor loading of the ith variable on the latent factor $\mathbf{f}_t$, and $\mathbf{u}_t$ is an idiosyncratic error vector. In the above model, $\mathbf{y}_t$ is the only observable variable, while $\mathbf{B}$ is a matrix of unknown parameters, and $(\mathbf{f}_t, \mathbf{u}_t)$ are latent random variables. Without loss of generality, we assume $\mathrm{E}(\mathbf{f}_t) = \mathrm{E}(\mathbf{u}_t) = \mathbf{0}$ and $\mathbf{f}_t$ and $\mathbf{u}_t$ are uncorrelated. Then, the model implied covariance structure is

$$\boldsymbol{\Sigma} = \mathbf{B}\mathrm{cov}(\mathbf{f}_t)\mathbf{B}' + \boldsymbol{\Sigma}_u,$$

where $\boldsymbol{\Sigma} = \mathrm{E}(\mathbf{y}_t\mathbf{y}_t')$ and $\boldsymbol{\Sigma}_u = \mathrm{E}(\mathbf{u}_t\mathbf{u}_t')$. Observe that $\mathbf{B}$ and $\mathbf{f}_t$ are not individually identifiable, since $\mathbf{B}\mathbf{f}_t = \mathbf{B}\mathbf{H}\mathbf{H}'\mathbf{f}_t$ for any orthogonal matrix $\mathbf{H}$. To this end, an identifiability condition is imposed:

$$\mathrm{cov}(\mathbf{f}_t) = \mathbf{I}_K \text{ and } \mathbf{B}'\boldsymbol{\Sigma}_u^{-1}\mathbf{B} \text{ is diagonal,} \tag{2}$$

which is a common assumption in the literature (Bai and Li, 2012; Bai and Liao, 2013).

Assume that we are only interested in a subset S among a total of p variables in model (1). We aim to obtain an efficient estimator of

$$\boldsymbol{\Sigma}_S = \mathbf{B}_S\mathbf{B}_S' + \boldsymbol{\Sigma}_{u,S},$$

the covariance matrix of the s variables in S, where $\mathbf{B}_S$ is the submatrix of $\mathbf{B}$ with row indices in S and $\boldsymbol{\Sigma}_{u,S}$ is the submatrix of $\boldsymbol{\Sigma}_u$ with row and column indices in S. As mentioned above, the existing literature uses the following conventional method:

- Method 1: Use solely the s variables in the set S to estimate common factors $\mathbf{f}_t$, the loading matrix $\mathbf{B}_S$, the idiosyncratic matrix $\boldsymbol{\Sigma}_{u,S}$, and the covariance matrix $\boldsymbol{\Sigma}_S$.

This idea is apparently strongly influenced by the nonparametric estimation of the covariance matrix and ignores a large portion of the available data in the other $p - s$ variables. An intuitively more efficient method is

- Method 2: Use all the p variables to obtain estimators of $\mathbf{f}_t$, the loading matrix $\mathbf{B}$, the idiosyncratic matrix $\mathbf{\Sigma}_u$, and the entire covariance matrix $\mathbf{\Sigma}$, and then restrict them to the variables of interest. This is the same as estimating $\mathbf{f}_t$ using all variables, and then estimating $\mathbf{B}_S$ and $\mathbf{\Sigma}_{u,S}$ based on the model (1) and the subset S with $\mathbf{f}_t$ being estimated (observed), and obtaining a plug-in estimator of $\mathbf{\Sigma}_S$.

We will show that Method 2 is more efficient than Method 1 in the estimation of $\mathbf{f}_t$ and $\mathbf{\Sigma}_S$ as more auxiliary data information is incorporated. By treating common factor as an unknown parameter, we calculate its Fisher information that grows with more data being utilized in Method 2. In this case, a more efficient factor estimate can be obtained, e.g., through weighted principal component (WPC) method (Bai and Liao, 2013). The advantage of factor estimation is further carried over to the estimation of $\mathbf{\Sigma}_S$ by Method 2 in terms of its convergence rate. Moreover, if the number of total variables is sufficiently large, Method 2 is proven to perform as well as an "oracle method", which observes all latent factors. This lends further support to our aforementioned claim of "blessing of dimensionality." Such a best possible rate improvement is new to the existing literature, and counted as another contribution of this paper. All these conclusions hold when the number of factors K is assumed to be fixed and known, while s, p and T all tend to infinity.

The idea of utilizing data as much as possible brings computational challenges. Fortunately, we observe that all the p variables are controlled by the same group of latent factors. Having said that, we can actually *split p variables* into smaller groups, and then utilize each group to estimate latent factors. The final factor estimate is obtained by averaging over these repeatedly estimated factors. Obviously, this divide-and-conquer algorithm can be implemented in a parallel computing environment, and thus produces factor estimators in a much more efficient way. On the other hand, our theory illustrates that this new method performs as well as the "pooled analysis", where we run the method over the whole dataset. Simulation studies further demonstrate the boosted computational speed and satisfactory statistical performance.

The rest of the paper is organized as follows. We compare the Fisher information of the factors by the two methods in Section 2. Section 3 describes the WPC method. As a main result, the convergence rates of different estimators of $\mathbf{\Sigma}_S$ are further compared in

Section 4 under various norms. Section 5 introduces the divide-and-conquer method for accelerating computation, while Section 6 presents all simulation results. Section 7 gives a microarray data example to illustrate our proposal. All technical proofs are delegated to the Appendix.

For any vector $\mathbf{a}$, let $\mathbf{a}_S$ denote a sub-vector of $\mathbf{a}$ with indices in S. Denote $\|\mathbf{a}\|$ the Euclidean norm of $\mathbf{a}$. For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, let $\mathbf{A}_{I,J}$ be the submatrix of $\mathbf{A}$ with row and column indices in I and J, respectively. We write $\mathbf{A}_S$ for $\mathbf{A}_{S,S}$ for simplicity. Let $\lambda_j(\mathbf{A})$ be the jth largest eigenvalue of $\mathbf{A}$. Denote $\|\mathbf{A}\| = \max\{|\lambda_1(\mathbf{A})|, |\lambda_d(\mathbf{A})|\}$ the operator norm of $\mathbf{A}$, $\|\mathbf{A}\|_{\max} = \max_{ij} |a_{ij}|$ the max-norm of $\mathbf{A}$, where a_{ij} is the (i,j)-th entry of $\mathbf{A}$, $\|\mathbf{A}\|_1 = \max_i \sum_{j=1}^d |a_{ij}|$ the L_1 norm of $\mathbf{A}$, $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}'\mathbf{A})}$ the Frobenius norm of $\mathbf{A}$, and $\|\mathbf{A}\|_{\mathbf{M}} = d^{-1/2}\|\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}\|_F$ the relative norm of $\mathbf{A}$ to $\mathbf{M}$, where the weight matrix $\mathbf{M}$ is assumed to be positive definite. For a non-square matrix $\mathbf{C}$, let $\mathbf{C}_S$ be the submatrix of $\mathbf{C}$ with row indices in S.

2 Fisher Information of Common Factor

In this section, we treat the vector of common factors as a fixed unknown parameter, and compute its Fisher information matrices based on Method 1 and Method 2. In the computation, the loading matrix $\mathbf{B}$ is treated as deterministic in Proposition 2. In Proposition 3, the Fisher information is computed for each given $\mathbf{B}$ and then averaged over $\mathbf{B}$ by regarding it as a realization of a chance process, which bypasses the block diagonal assumption needed without taking average over $\mathbf{B}$. In other sections, we adopt the convention regarding the factors as random and $\mathbf{B}$ as fixed. We start by calculating the Fisher information of $\boldsymbol{\theta}_t := \mathbf{B}\mathbf{f}_t$, which serves as an intermediate step in obtaining that for $\mathbf{f}_t$. For simplicity of notation, time t is suppressed in $(\mathbf{y}_t, \mathbf{f}_t, \mathbf{u}_t, \boldsymbol{\theta}_t)$ so that it becomes $(\mathbf{y}, \mathbf{f}, \mathbf{u}, \boldsymbol{\theta})$ in this section.

Given a general density function of $\mathbf{y}$, denoted as $h(\mathbf{y};\boldsymbol{\theta})$, the Fisher information of $\boldsymbol{\theta}$ contained in full data is given by

$$I_p(\boldsymbol{\theta}) = \mathrm{E}\left[\left(\frac{\partial \log h(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log h(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right].$$

When only data in S is used, the Fisher information of $\boldsymbol{\theta}_S$ is given by

$$I_S(\boldsymbol{\theta}_S) = \mathrm{E}\left[\left(\frac{\partial \log h_S(\mathbf{y}_S;\boldsymbol{\theta}_S)}{\partial \boldsymbol{\theta}_S}\right)\left(\frac{\partial \log h_S(\mathbf{y}_S;\boldsymbol{\theta}_S)}{\partial \boldsymbol{\theta}_S}\right)'\right],$$

where h_S is the marginal density of $\mathbf{y}_S$ for the target set of variable S. Our first proposition shows that $\{I_p(\boldsymbol{\theta})\}_S$, the submatrix of $I_p(\boldsymbol{\theta})$ restricted on S, dominates $I_S(\boldsymbol{\theta}_S)$ under a mild condition.

Proposition 1. *If $h(\mathbf{y};\boldsymbol{\theta}) = h(\mathbf{y}-\boldsymbol{\theta})$ and the density function $h(\mathbf{y}-\boldsymbol{\theta})$ satisfies the following regularity condition:*

$$\nabla_{\mathbf{y}_S} \int h(\mathbf{y}_S - \boldsymbol{\theta}_S, \mathbf{y}_{S^c} - \boldsymbol{\theta}_{S^c}) \mathrm{d}\mathbf{y}_{S^c} = \int \nabla_{\mathbf{y}_S} h(\mathbf{y}_S - \boldsymbol{\theta}_S, \mathbf{y}_{S^c} - \boldsymbol{\theta}_{S^c}) \mathrm{d}\mathbf{y}_{S^c}, \tag{3}$$

then $\{I_p(\boldsymbol{\theta})\}_S \succeq I_S(\boldsymbol{\theta}_S)$ in the sense that $\{I_p(\boldsymbol{\theta})\}_S - I_S(\boldsymbol{\theta}_S)$ is positive semi-definite.

The regularity condition (3) is fairly mild, as illustrated in the following examples.

Example 1. In model 1, if $\mathbf{u}_S$ and $\mathbf{u}_{S^c}$ are independent, then (3) holds.

Example 2. If $\mathbf{y}$ follows an elliptical distribution that

$$h(\mathbf{y};\boldsymbol{\theta}) \propto g((\mathbf{y}-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\theta})),$$

where the mapping function $g(t) : [0,\infty) \to [0,\infty)$ satisfies that $|g'(t)| \leq cg(t)$ for some positive constant c, and $\mathrm{E}|\mathbf{y}| < \infty$, then (3) holds. Example 2 includes some commonly used multivariate distributions as its special cases, e.g. the multivariate normal distribution and the multivariate t-distribution with degrees of freedom greater than 1. The proof is given in the Appendix Section A.2.

We next compute the Fisher information of $\mathbf{f}$ based on the full data set, denoted as $I(\mathbf{f})$, and the partial data set restricted on S, denoted as $I_S(\mathbf{f})$. This can be done easily by noting that $I(\mathbf{f}) = \mathbf{B}'I_p(\boldsymbol{\theta})\mathbf{B}$. Indeed, the WPC estimators used in Methods 1 and 2 achieve such efficiency since their asymptotic variances are proven to be the inverse of $I(\mathbf{f})$ and $I_S(\mathbf{f})$, respectively; see Remark 1.

Proposition 2 shows that $I(\mathbf{f})$ dominates $I_S(\mathbf{f})$, if $I_p(\boldsymbol{\theta})$ is block-diagonal, i.e., $\{I_p(\boldsymbol{\theta})\}_{S,S^c} = \mathbf{0}$. Hence, common factors can be estimated more efficiently using additional data $\mathbf{y}_{S^c}$. The above block-diagonal condition implies that the idiosyncratic error of additional variables cannot be confounded with that of the variables-of-interest. For example, if $\mathbf{u}$ is normal, then $\{I_p(\boldsymbol{\theta})\}_{S,S^c} = \mathbf{0}$ indeed requires that $\mathbf{u}_S$ is independent of $\mathbf{u}_{S^c}$.

Proposition 2. *Under condition (3), if* $\{I_p(\boldsymbol{\theta})\}_{S,S^c} = \mathbf{0}$, $I(\mathbf{f}) \succeq I_S(\mathbf{f})$.

So far we treat $\mathbf{B}$ as being deterministic. Rather, Proposition 3 regards $\{\mathbf{b}_i\}$ as a realization of a chance process. Under this assumption, the expectation of $I(\mathbf{f})$ over $\mathbf{B}$ is shown to always dominate that of $I_S(\mathbf{f})$. In other words, we can claim that averaging over loading matrices, a larger dataset contains more information about the unknown factors.

Proposition 3. *If* $\{\mathbf{b}_i\}_{i=1}^p$ *are i.i.d. random loadings with* $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$ *and (3) holds, then* $\mathrm{E}[I(\mathbf{f})] \succeq \mathrm{E}[I_S(\mathbf{f})]$, *where the expectation is taken with respect to the distribution of* $\mathbf{B}$.

3 Efficient Estimation of Common Factor

In this section, we construct an efficient estimator of the common factors by showing that its asymptotic variance is exactly the inverse of its Fisher information. This together with the arguments in Section 2 enables us to draw a conclusion that using more data results in a more efficient factor estimator with a smaller asymptotic variance.

From a least-squares perspective, when the loading matrix $\mathbf{B}$ is known, $\mathbf{f}_t$ can be estimated by the weighted least-squares: $\mathrm{argmin}_{\mathbf{f}_t \in \mathbb{R}^K} \sum_{t=1}^T (\mathbf{y}_t - \mathbf{B}\mathbf{f}_t)'\boldsymbol{\Sigma}_u^{-1}(\mathbf{y}_t - \mathbf{B}\mathbf{f}_t)$. In the high-dimensional setting ($p \gg T$), we assume $\boldsymbol{\Sigma}_u$ is a sparse matrix and define its sparsity measurement as

$$m_p = \max_{i \le p} \sum_{j \ne i} I(\sigma_{u,ij} \ne 0), \text{ where } \sigma_{u,ij} \text{ is the } (i,j)\text{-th entry of } \boldsymbol{\Sigma}_u. \tag{4}$$

In particular, we assume the following sparsity condition

$$m_p = o\left(\min\left\{\frac{1}{p^{1/4}}\sqrt{\frac{T}{\log p}}, p^{1/4}\right\}\right) \text{ and } \sum_{i=1}^p \sum_{j \ne i} I(\sigma_{u,ij} \ne 0) = O(p). \tag{5}$$

Now, we propose to solve the following constrained weighted least-squares problem:

$$\begin{aligned} (\widehat{\mathbf{B}}, \widehat{\mathbf{f}}_1, \ldots, \widehat{\mathbf{f}}_T) &= \mathop{\mathrm{argmin}}_{\mathbf{B}, \mathbf{f}_t} \sum_{t=1}^T (\mathbf{y}_t - \mathbf{B}\mathbf{f}_t)'\tilde{\boldsymbol{\Sigma}}_u^{-1}(\mathbf{y}_t - \mathbf{B}\mathbf{f}_t), \\ &\text{subject to } \frac{1}{T}\sum_{t=1}^T \mathbf{f}_t\mathbf{f}_t' = \mathbf{I}_K;\ \mathbf{B}'\tilde{\boldsymbol{\Sigma}}_u^{-1}\mathbf{B} \text{ is diagonal,} \end{aligned} \tag{6}$$

where $\tilde{\boldsymbol{\Sigma}}_u$ is a regularized estimator of $\boldsymbol{\Sigma}_u$ to be discussed later. The above constraint is a sample analog of the identifiability condition (2). The involvement of the weight $\tilde{\boldsymbol{\Sigma}}_u^{-1}$ is to account for the heterogeneity among the data and leads to more efficient estimation of $(\mathbf{B}, \mathbf{f}_t)$ (Choi, 2012; Bai and Liao, 2013).

Indeed, an initial estimator $\tilde{\boldsymbol{\Sigma}}_u$ of the idiosyncratic matrix $\boldsymbol{\Sigma}_u$ is needed for solving the constrained weighted least-squares problem. We propose to obtain such an estimator by the following procedure, which is in the same spirit as the estimation of the idiosyncratic matrix in the POET method (Fan et al., 2013). Let $\mathbf{S}_y = T^{-1}\sum_{t=1}^{T}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})'$ be the sample covariance of $\mathbf{y}$ and $\{(\lambda_i, \boldsymbol{\zeta}_i)\}_{i=1}^{p}$ be eigen-pairs of $\mathbf{S}_y$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Denote $\mathbf{R} = \mathbf{S}_y - \sum_{i=1}^{K} \lambda_i \boldsymbol{\zeta}_i \boldsymbol{\zeta}_i'$. We estimate $\boldsymbol{\Sigma}_u$ by $\tilde{\boldsymbol{\Sigma}}_u$, whose (i, j)-th entry

$$\widehat{\sigma}_{u,ij} = \begin{cases} r_{ii}, & \text{for } i = j, \\ s_{ij}(r_{ij}), & \text{for } i \neq j, \end{cases} \quad \text{where } \mathbf{R} = (r_{ij}),$$

$s_{ij}(r_{ij})$ is a general entry-wise thresholding function (Antoniadis and Fan, 2001) such that $s_{ij}(z) = 0$ if $|z| \leq \tau_{ij}$ and $|s_{ij}(z) - z| \leq \tau_{ij}$ for $|z| > \tau_{ij}$. In our paper, we choose hard-thresholding even though SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) are also applicable. We specify the entry-wise thresholding level as

$$\tau_{ij}(p) = C\sqrt{r_{ii}r_{jj}}\,\omega(p), \text{ where } \omega(p) = \sqrt{\frac{\log p}{T}} + \frac{1}{\sqrt{p}}, \tag{7}$$

and C is a constant chosen by cross-validation. The thresholding parameter $C\omega(p)$ is applied to the correlation matrix. This is similar to the adaptive thresholding estimator for a general covariance matrix (Rothman et al., 2009), where the entry-wise thresholding level depends on p.

With $\tilde{\boldsymbol{\Sigma}}_u$ being the thresholding estimator described above, the constrained weighted least-squares problem (6) can be solved by the weighted principal component (WPC) method. The solution is given by

$$\widehat{\mathbf{F}} = (\widehat{\mathbf{f}}_1, \ldots, \widehat{\mathbf{f}}_T)' \text{ and } \widehat{\mathbf{B}} = T^{-1}\mathbf{Y}\widehat{\mathbf{F}}, \tag{8}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$ and the columns of $\widehat{\mathbf{F}}$ are the eigenvectors corresponding to the largest K eigenvalues of the $T \times T$ matrix $\sqrt{T}\mathbf{Y}'\tilde{\boldsymbol{\Sigma}}_u^{-1}\mathbf{Y}$ (Bai and Liao, 2013).

In the following, we give a result showing that the WPC estimator is asymptotically efficient. Indeed, Bai and Liao (2013) derive the asymptotic normality of $\widehat{\mathbf{f}}_t$ under the following conditions:

(i) All eigenvalues of $\mathbf{B}'\mathbf{B}/p$ are bounded away from zero and infinity as $p \to \infty$;

(ii) There exists a $K \times K$ diagonal matrix $\mathbf{Q}$ such that $\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}/p \to \mathbf{Q}$. In addition, the diagonal elements of $\mathbf{Q}$ are distinct and bounded away from infinity.

(iii) For each fixed $t \leq T$, $(\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B})^{-1/2}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_K)$, as $p \to \infty$,

together with the sparsity assumption (5), and some additional regularity conditions given in Section A.1. When $\sqrt{p}\log p = o(T)$, it is shown that

$$\sqrt{p}(\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t) \xrightarrow{D} N(\mathbf{0}, \mathbf{Q}^{-1}), \tag{9}$$

where $\mathbf{H}$ is a specific rotation matrix given by

$$\mathbf{H} = \widehat{\mathbf{V}}^{-1}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{B}/T, \tag{10}$$

and $\widehat{\mathbf{V}}$ is a $K \times K$ diagonal matrix of the largest K eigenvalues of $\mathbf{Y}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{Y}/T$. The rotation matrix $\mathbf{H}$ is introduced here so that $\mathbf{H}\mathbf{f}_t$ is an identifiable quantity from the data. See more discussion about the identifiability in Remark 2.

Condition (i) is a "pervasive condition" requiring that the common factors affect a non-negligible fraction of subjects. This is a common assumption for the principal components based methods (Fan et al., 2011; Bai and Liao, 2013). In condition (ii), $\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B}$ is indeed the Fisher information (under Gaussian errors) contained in p variables, while the limit $\mathbf{Q}$ can be viewed as an average information for each variable. Hence, the asymptotic normality in (9) shows that $\widehat{\mathbf{f}}_t$ is efficient as its asymptotic variance attains the inverse of the (averaged) Fisher information.

Remark 1. The results in Section 2 together with (9) imply that Method 2 is in general better than Method 1 in the estimation of common factors. To explain why, we consider two different cases here. When p is an order of magnitude larger than s, where s is the number of variables of interest. Method 2 produces a better estimator of factors with a faster convergence rate. Even when p and s diverge at the same speed, the factor estimator based on Method 2 is shown to possess a smaller asymptotic variance, as long as $\mathbf{\Sigma}_{u,S,S^c} = \mathbf{0}$. Recall that $\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{B} = I(\mathbf{f})$ and $\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S = I_S(\mathbf{f})$ under Gaussian errors, and they also correspond to the inverse of the asymptotic variance given by Methods 1 and 2, respectively.

Then, Proposition 2 implies that Method 2 has a smaller asymptotic variance, if $\boldsymbol{\Sigma}_{u,S,S^c} = \mathbf{0}$. Alternatively, if $\mathbf{B}$ is treated as being random, Proposition 3 immediately implies that $E(\mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S) \succeq E(\mathbf{B}'\boldsymbol{\Sigma}^{-1}\mathbf{B})$. Therefore, even without the block diagonal assumption, Method 2 produces a more efficient factor estimate on average.

4 Covariance Matrix Estimation

One primary goal in this paper is to obtain an accurate estimator of the covariance matrix $\boldsymbol{\Sigma}_S = \mathrm{E}(\mathbf{y}_S\mathbf{y}_S')$ for the variables-of-interest. In this section, we compare three different estimation methods, namely Methods 1, 2 and Oracle Method, in terms of their rates of convergence (under various norms). Obviously, these rates depend on how accurately the realized factors are estimated as demonstrated later.

Below we describe these three methods in full details.

- Method 1:

  i. Use solely the data in the subset S to obtain estimators of the realized factors $\widehat{\mathbf{F}}^{(1)}$ and the loading matrix $\widehat{\mathbf{B}}_1 = T^{-1}\mathbf{Y}_S\widehat{\mathbf{F}}^{(1)}$ based on (8);

  ii. Let $(\widehat{\mathbf{f}}_t^{(1)})'$ be the t-th row of $\widehat{\mathbf{F}}^{(1)}$, $(\widehat{\mathbf{b}}_i^{(1)})'$ be the i-th row of $\widehat{\mathbf{B}}_1$, $\widehat{u}_{it} = y_{it} - (\widehat{\mathbf{b}}_i^{(1)})'\widehat{\mathbf{f}}_t^{(1)}$, and $\widehat{\sigma}_{ij} = \frac{1}{T}\sum_{t=1}^T \widehat{u}_{it}\widehat{u}_{jt}$. The (i,j)-th entry of the idiosyncratic matrix estimator $\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)}$ of $\boldsymbol{\Sigma}_{u,S}$ is given by thresholding $\widehat{\sigma}_{ij}$ at the level of $C\widehat{\theta}_{ij}^{1/2}\omega(s)$, where $\omega(s)$ is defined in (7) and $\widehat{\theta}_{ij} = \frac{1}{T}\sum_{t=1}^T(\widehat{u}_{it}\widehat{u}_{jt} - \widehat{\sigma}_{ij})^2$;

  iii. The final estimator is given by $\widehat{\boldsymbol{\Sigma}}_S^{(1)} = \widehat{\mathbf{B}}_1\widehat{\mathbf{B}}_1' + \widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)}$.

- Method 2:

  i. Use all p variables to obtain the estimate $\widehat{\mathbf{F}}^{(2)}$ as given in (8) for the realized factors and then estimate the loading $\mathbf{B}_S$ by $\widehat{\mathbf{B}}_2 = T^{-1}\mathbf{Y}_S\widehat{\mathbf{F}}^{(2)}$;

  ii. Follow the same procedure as in Method 1 to obtain the estimator $\widehat{\boldsymbol{\Sigma}}_{u,S}^{(2)}$ but based on $\widehat{\mathbf{F}}^{(2)}$ and $\widehat{\mathbf{B}}_2$;

  iii. The final estimator is given by $\widehat{\boldsymbol{\Sigma}}_S^{(2)} = \widehat{\mathbf{B}}_2\widehat{\mathbf{B}}_2' + \widehat{\boldsymbol{\Sigma}}_{u,S}^{(2)}$.

- Oracle Method:

i. Estimate the loading by $\widehat{\mathbf{B}}_o = T^{-1}\mathbf{Y}_S\mathbf{F}$, where $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)'$ are the true factors.

ii. The idiosyncratic matrix estimator $\widehat{\boldsymbol{\Sigma}}^o_{u,S}$ is given by the same procedure as in Method 1, with $\widehat{\mathbf{b}}_i^{(1)}$ and $\widehat{\mathbf{f}}_t^{(1)}$ being replaced by $\widehat{\mathbf{b}}_i^o$ and $\mathbf{f}_t$, respectively.

iii. The final estimator is given by $\widehat{\boldsymbol{\Sigma}}^o_S = \widehat{\mathbf{B}}_o\widehat{\mathbf{B}}_o' + \widehat{\boldsymbol{\Sigma}}^o_{u,S}$.

Theorem 1 depicts the estimation accuracy of $\boldsymbol{\Sigma}_S$ by the above three methods with respect to the following measurements:

$$\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|_{\boldsymbol{\Sigma}_S}, \quad \|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|_{\max}, \quad \|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\|,$$

where $\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|_{\boldsymbol{\Sigma}_S} = p^{-1/2}\|\boldsymbol{\Sigma}_S^{-1/2}\widehat{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2} - \mathbf{I}_S\|_F$ is a norm of the relative errors. Note that the results of Fan et al. (2013) can not be directly used here since we employ the *weighted* principal component analysis to estimate the unobserved factors. This is expected to be more accurate than the ordinary principal component analysis, as shown in Bai and Liao (2013). Indeed, the technical proofs for our results are technically more involved than those in Fan et al. (2013).

We assume that s is much less than p, i.e., $s = o(p)$, but both tend to infinity. Under the pervasive condition (i), $\|\boldsymbol{\Sigma}_S\| \geq cs$ and therefore diverges. For this reason, we consider the relative norm $\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|_{\boldsymbol{\Sigma}_S}$, instead of $\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|$, and the operator norm $\|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\|$ for estimating the inverse. In addition, we consider another element-wise max norm $\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|_{\max}$. We show that if p is large with respect to s and T, Method 2 performs as well as the Oracle Method, both of which outperform Method 1. As a consequence, even if we are only interested in the covariance matrix of a small subset of variables, we should use all the data to estimate the common factors, which ultimately improves the estimation of $\boldsymbol{\Sigma}_S$. In particular, we are able to specify an explicit regime of (s, p) under which the improvements are substantial. However, when $s \asymp p$, i.e. they are in the same order, using more data does not show as dramatic improvements for estimating $\boldsymbol{\Sigma}_S$. This is expected and will be clearly seen in the simulation section.

Before stating Theorem 1, we need a few preliminary results: Lemmas 1 – 3. Specifically, Lemma 1 presents the uniform convergence rates of the factor estimates by Methods 1 and 2. Based on that, Lemmas 2 and 3 further derive the estimation accuracy of factor loadings and idiosyncratic matrix by the three methods, respectively. These results together lead to the estimation error rates of $\mathbf{\Sigma}_S$ in Theorem 1 w.r.t. three measures defined above. Additional Lemmas supporting the proof are given in Appendix. Again, these kinds of results can not be obtained directly from Fan et al. (2013) due to our use of WPC.

Lemma 1. *Suppose that conditions (i), (ii), the sparsity condition (5), and additional regularity conditions (iv)-(vii) in Section A.1 hold for both s and p. If $\sqrt{p}\log p = o(T)$ and $T = o(s^2)$, then we have*

$$\max_{t\leq T}\|\widehat{\mathbf{f}}_t^{(1)} - \mathbf{H}_1\mathbf{f}_t\| = O_P\left(\frac{1}{\sqrt{T}} + \frac{T^{1/4}}{\sqrt{s}}\right) \quad \textit{and} \quad \max_{t\leq T}\|\widehat{\mathbf{f}}_t^{(2)} - \mathbf{H}_2\mathbf{f}_t\| = O_P\left(\frac{1}{\sqrt{T}} + \frac{T^{1/4}}{\sqrt{p}}\right),$$

where $\mathbf{H}_1 = \widehat{\mathbf{V}}_1^{-1}\widehat{\mathbf{F}}^{(1)\prime}\mathbf{F}\mathbf{B}_S'\tilde{\mathbf{\Sigma}}_{u,S}^{-1}\mathbf{B}_S/T$, $\mathbf{H}_2 = \widehat{\mathbf{V}}_2^{-1}\widehat{\mathbf{F}}^{(2)\prime}\mathbf{F}\mathbf{B}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{B}/T$, $\widehat{\mathbf{V}}_1$ *is the diagonal matrix of the largest K eigenvalues of* $\mathbf{Y}_S'\tilde{\mathbf{\Sigma}}_{u,S}^{-1}\mathbf{Y}_S/T$ *and* $\widehat{\mathbf{V}}_2$ *is the diagonal matrix of the largest K eigenvalues of* $\mathbf{Y}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{Y}/T$.

Remark 2. $\mathbf{H}_1$ and $\mathbf{H}_2$ correspond to the rotation matrix $\mathbf{H}$ defined in (10) using Methods 1 and 2, respectively. Recall that $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_T)'$, then $\mathbf{H}\mathbf{f}_t = T^{-1}\widehat{\mathbf{V}}^{-1}\widehat{\mathbf{F}}(\mathbf{B}\mathbf{f}_1, \ldots, \mathbf{B}\mathbf{f}_T)'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{B}\mathbf{f}_t$. Note that $\mathbf{H}\mathbf{f}_t$ only depends on quantities $\mathbf{V}^{-1}\widehat{\mathbf{F}}$, $\tilde{\mathbf{\Sigma}}_u^{-1}$ and the *identifiable* component $\{\mathbf{B}\mathbf{f}_t\}_{t=1}^T$. Therefore, there is no identifiability issue regarding $\mathbf{H}\mathbf{f}_t$. In other words, even though $\mathbf{f}_t$ itself may not be identifiable, an identifiable rotation of $\mathbf{f}_t$ can be consistently estimated by $\widehat{\mathbf{f}}_t$.

Lemma 1 implies that Method 2 produces a better factor estimate if

$$0.5 < \gamma_s < 1.5 \leq \gamma_p < 2,$$

by representing s and p as $s \asymp T^{\gamma_s}$ and $p \asymp T^{\gamma_p}$.

It is not surprising that the estimation accuracy of loading matrix also varies among these three methods as shown in Lemma 2 below.

Lemma 2. *Under conditions of Lemma 1,*

$$\max_{i\leq s}\|\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}_1\mathbf{b}_i\| = O_P(w_1), \ \textit{where } w_1 := \frac{1}{\sqrt{s}} + \sqrt{\frac{\log s}{T}},$$

$$\max_{i\leq s}\|\widehat{\mathbf{b}}_i^{(2)} - \mathbf{H}_2\mathbf{b}_i\| = O_P(w_2), \ \textit{where } w_2 := \frac{1}{\sqrt{p}} + \sqrt{\frac{\log s}{T}},$$

$$\max_{i\leq s}\|\widehat{\mathbf{b}}_i^{o} - \mathbf{b}_i\| = O_P(w_o), \ \textit{where } w_o := \sqrt{\frac{\log s}{T}}.$$

Similarly, Lemma 2 indicates that Method 2 performs as well as the Oracle Method, both of which are better than Method 1, i.e., $w_2 = w_o < w_1$, if

$$0.5 < \gamma_s < 1 \leq \gamma_p < 2,$$

by representing s and p in the order of T as above. We remark that the extra terms $1/\sqrt{s}$ and $1/\sqrt{p}$ in w_1 and w_2 (in comparison with the oracle rate w_o) are due to the factor estimation. Another preliminary result regarding the estimation of the identifiable component $\mathbf{b}_i'\mathbf{f}_t$ is given in Lemma A.1.

Similar insights can be delivered from Lemma 3 on the estimation of $\mathbf{\Sigma}_{u,S}$.

Lemma 3. *Under conditions of Lemma 1, it holds that*

$$\begin{aligned}
\|\widehat{\mathbf{\Sigma}}_{u,S}^{(1)} - \mathbf{\Sigma}_{u,S}\| = O_P\left(m_s w_1\right) = \|(\widehat{\mathbf{\Sigma}}_{u,S}^{(1)})^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\|,\\
\|\widehat{\mathbf{\Sigma}}_{u,S}^{(2)} - \mathbf{\Sigma}_{u,S}\| = O_P\left(m_s w_2\right) = \|(\widehat{\mathbf{\Sigma}}_{u,S}^{(2)})^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\|,\\
\|\widehat{\mathbf{\Sigma}}_{u,S}^{o} - \mathbf{\Sigma}_{u,S}\| = O_P\left(m_s w_o\right) = \|(\widehat{\mathbf{\Sigma}}_{u,S}^{o})^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\|,
\end{aligned}$$

where m_s is defined as in (4) with p being replaced by s.

Now, we are ready to state our main result on the estimation of $\mathbf{\Sigma}_S$ based on the above preliminary results. From Theorem 1, it is easily seen that the comparison of the estimation accuracy of $\mathbf{\Sigma}_S$ among three methods is solely determined by the relative magnitude of w_o, w_1 and w_2. Therefore, we should use additional variables to estimate the factors if p is much larger than s in the sense that $T/\log s = O(p)$ and $s \log s = o(T)$ (implying $w_2 = w_o < w_1$).

Theorem 1. *Under conditions of Lemma 1 , it holds that*

(1) For the relative norm, $\|\widehat{\mathbf{\Sigma}}_S^{(1)} - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S} = O_P\left(\sqrt{s}w_1^2 + m_s w_1\right)$, $\|\widehat{\mathbf{\Sigma}}_S^{(2)} - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S} = O_P\left(\sqrt{s}w_2^2 + m_s w_2\right)$, *and* $\|\widehat{\mathbf{\Sigma}}_S^{o} - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S} = O_P\left(\sqrt{s}w_o^2 + m_s w_o\right)$.

(2) For the max-norm, $\|\widehat{\mathbf{\Sigma}}_S^{(1)} - \mathbf{\Sigma}_S\|_{\max} = O_P\left(w_1\right)$, $\|\widehat{\mathbf{\Sigma}}_S^{(2)} - \mathbf{\Sigma}_S\|_{\max} = O_P\left(w_2\right)$, *and* $\|\widehat{\mathbf{\Sigma}}_S^{o} - \mathbf{\Sigma}_S\|_{\max} = O_P\left(w_o\right)$.

(3) For the operator norm of the inverse matrix, $\|(\widehat{\mathbf{\Sigma}}_S^{(1)})^{-1} - \mathbf{\Sigma}_S^{-1}\| = O_P\left(m_s w_1\right)$, $\|(\widehat{\mathbf{\Sigma}}_S^{(2)})^{-1} - \mathbf{\Sigma}_S^{-1}\| = O_P\left(m_s w_2\right)$ *and* $\|(\widehat{\mathbf{\Sigma}}_S^{o})^{-1} - \mathbf{\Sigma}_S^{-1}\| = O_P\left(m_s w_o\right)$.

Remark 3. So far, we assumed that the number of factors K is fixed and known. A data driven choice of K has been extensively studied in the econometrics literature, e.g., by Bai and Ng (2002), Kapetanios (2010). To estimate K, we can adopt the method by Bai and Ng (2002) and propose a consistent estimator of K (by allowing $p, T \to \infty$) as follows

$$\widehat{K} = \underset{0 \le k \le N}{\operatorname{argmin}} \log \left\{ \frac{1}{pT} \|\mathbf{Y} - T^{-1}\mathbf{Y}\widehat{\mathbf{F}}_k \widehat{\mathbf{F}}_k'\|_F^2 \right\} + k g(p, T),$$

where N is a predefined upper bound, $\widehat{\mathbf{F}}_k$ is a $T \times k$ matrix whose columns are $\sqrt{T}$ times the eigenvectors corresponding to the largest k eigenvalues of $\mathbf{Y}'\mathbf{Y}$, and $g(p, T)$ is a penalty function. Two examples suggested by Bai and Ng (2002) are

$$g(T, p) = \frac{p+T}{pT} \log\left(\frac{pT}{p+T}\right) \quad \text{or} \quad g(T, p) = \frac{p+T}{pT} \log\left(\min\{p, T\}\right).$$

Under our assumptions (i)-(x), all conditions required by theorem 2 of Bai and Ng (2002) hold. Hence, their theorem implies that $P(\widehat{K} = K) \to 1$. Then, conditioning on the event that $\{\widehat{K} = K\}$, our theorem 1 still holds by replacing K with $\widehat{K}$. Other effective methods for selecting the number of factors include the eigen ratio method in Lam and Yao (2012) and Ahn and Horenstein (2013).

Remark 4. When K grows with p and T, Fan et al. (2013) gives the explicit dependence of the convergence rates on K for their proposed POET estimator. By adopting their technique, we can obtain the following results:
(1) $\|\widehat{\mathbf{\Sigma}}_S^{(1)} - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S} = O_P\left(K\sqrt{s}w_1^2 + K^3 m_s w_1\right)$, $\|\widehat{\mathbf{\Sigma}}_S^{(2)} - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S} = O_P\left(K\sqrt{s}w_2^2 + K^3 m_s w_2\right)$, $\|\widehat{\mathbf{\Sigma}}_S^{o} - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S} = O_P\left(K\sqrt{s}w_o^2 + K^3 m_s w_o\right)$;
(2) $\|\widehat{\mathbf{\Sigma}}_S^{(1)} - \mathbf{\Sigma}_S\|_{\max} = O_P\left(K^3 w_1\right)$, $\|\widehat{\mathbf{\Sigma}}_S^{(2)} - \mathbf{\Sigma}_S\|_{\max} = O_P\left(K^3 w_2\right)$, $\|\widehat{\mathbf{\Sigma}}_S^{o} - \mathbf{\Sigma}_S\|_{\max} = O_P\left(K^3 w_o\right)$;
(3) $\|(\widehat{\mathbf{\Sigma}}_S^{(1)})^{-1} - \mathbf{\Sigma}_S^{-1}\| = O_P\left(K^3 m_s w_1\right)$, $\|(\widehat{\mathbf{\Sigma}}_S^{(2)})^{-1} - \mathbf{\Sigma}_S^{-1}\| = O_P\left(K^3 m_s w_2\right)$, $\|(\widehat{\mathbf{\Sigma}}_S^{o})^{-1} - \mathbf{\Sigma}_S^{-1}\| = O_P\left(K^3 m_s w_o\right)$.
Again, the rate difference among three types of estimators only depends on w_o, w_1 and w_2. Therefore, the same conclusion (when p is much larger than s, using additional variables improves the estimation of $\mathbf{\Sigma}_S$) can still be made even if K diverges. As long as K diverges in the rate that $K = o(\min\{1/(\sqrt{s}w_1^2), 1/(m_s w_1)^{1/3}\})$, $K = o(1/w_1^{1/3})$ or $K = o(1/(m_s w_1)^{1/3})$, the same blessing of dimensionality phenomena persist in terms of estimation consistency in relative norm, max norm, or operator norm of the inverse, respectively.

5 Divide-and-Conquer Computing Method

As discussed previously, we prefer utilizing auxiliary data information as much as possible even we are only interested in the covariance matrix of some particular set of variables. But this can bring up heavy computational burden. This concern motivates a simple divide-and-conquer scheme that *splits all p variables in* $\mathbf{Y}$. Without loss of generality, assume that p rows of matrix $\mathbf{Y}$ can be evenly divided into M groups with p/M variables in each group. The s variables of interest can possibly be assigned to different groups.

Divide-and-Conquer Computation Scheme

1. In the mth group, obtain the initial estimator $\tilde{\mathbf{\Sigma}}_{u,m}$ by using the adaptive thresholding method as described in Section 3 based on the data in the mth group only.

2. Denote $\mathbf{Y}_m$ as the data vector corresponding to the variables in the mth group and let $\widehat{\mathbf{F}}_m = (\widehat{\mathbf{f}}_{m,1}, \ldots, \widehat{\mathbf{f}}_{m,T})'$, where its columns are the eigenvectors corresponding to the largest K eigenvalues of the $T \times T$ matrix $\sqrt{T}\mathbf{Y}_m'\tilde{\mathbf{\Sigma}}_{u,m}^{-1}\mathbf{Y}_m$. The computation in the above two steps can be done in a parallel manner.

3. Average $\{\widehat{\mathbf{f}}_{m,t}\}_{m=1}^M$ to obtain a single estimator of $\mathbf{f}_t$ as

$$\bar{\mathbf{f}}_t = \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbf{f}}_{m,t}.$$

   The loading matrix estimate is given by $\bar{\mathbf{B}}_S = T^{-1}\mathbf{Y}_S\bar{\mathbf{F}}$, where $\bar{\mathbf{F}} = (\bar{\mathbf{f}}_1, \ldots, \bar{\mathbf{f}}_T)'$.

4. The idiosyncratic matrix is estimated as follows. Let $\bar{\mathbf{f}}_t'$ be the t-th row of $\bar{\mathbf{F}}$ and $\bar{\mathbf{b}}_i'$ be the ith row of $\bar{\mathbf{B}}_S$. Let $\widehat{u}_{it} = y_{it} - \bar{\mathbf{b}}_i'\bar{\mathbf{f}}_t$, $\widehat{\sigma}_{ij} = T^{-1}\sum_{t=1}^T \widehat{u}_{it}\widehat{u}_{jt}$, and $\widehat{\theta}_{ij} = T^{-1}\sum_{t=1}^T(\widehat{u}_{it}\widehat{u}_{jt} - \widehat{\sigma}_{ij})^2$. The (i,j)-th entry of $\bar{\mathbf{\Sigma}}_{u,S}$ is given by thresholding $\widehat{\sigma}_{ij}$ at the level of $C\widehat{\theta}_{ij}^{1/2}\omega(s)$, where $\omega(s)$ is defined as in (7) with p replaced by s.

5. The final estimator of the covariance matrix is given by

$$\bar{\mathbf{\Sigma}}_S = \bar{\mathbf{B}}_S\bar{\mathbf{B}}_S' + \bar{\mathbf{\Sigma}}_{u,S}.$$

We show that, if M is fixed,

$$\|\bar{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S} = O_P\left(\sqrt{s}w_2^2 + m_s w_2\right),$$

$$\|\bar{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|_{\max} = O_P\left(w_2\right),$$
$$\|(\bar{\mathbf{\Sigma}}_S)^{-1} - \mathbf{\Sigma}_S^{-1}\| = O_P\left(m_s w_2\right).$$

These rates match the rates of $\widehat{\mathbf{\Sigma}}_S^{(2)}$ attained by Method 2, where all p variables are pooled together for the analysis. The proof is given in Appendix A.3. The simulation results in Section 6 further demonstrate that without sacrificing the estimation accuracy, the divide-and-conquer method runs much faster than Method 2. Therefore, the divide-and-conquer method is practically useful when dealing with massive dataset.

The main computational cost of our method comes from taking the inverse of $\tilde{\mathbf{\Sigma}}_u$. For our Method 2, where all p variables are pooled together for the analysis, the computational complexity of the inversion is $O(p^3)$. On the other hand, for the divide-and-conquer method, the corresponding estimator $\tilde{\mathbf{\Sigma}}_{u,m}$ in the m-th group only needs a computational cost of $O((p/M)^3)$ to be inverted. Then, the total computation complexity is $O(p^3/M^2)$. Hence, the computational speed can be boosted by M^2-fold. Such a computational acceleration can also be observed from simulation study results in Figure 1(d). Other operations like the eigen-decomposition on the $T\times T$ matrix $\sqrt{T}\mathbf{Y}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{Y}$ do not have dominating computational cost, as we assume that p is much larger than T. When M grows too fast, the divide-and-conquer method may lose estimation efficiency compared with the pooled analysis (Method 2). However, considering its boost of computation, the divide-and-conquer method is practically useful when dealing with massive dataset.

6 Simulations

We use simulated examples to compare the statistical performances of Methods 1, 2 and the Oracle Method. We fix the number of factors $K = 3$ and repeat 100 simulations for each combination of (s, p, T). The loading $\mathbf{b}_i$, the factor $\mathbf{f}_t$ and the idiosyncratic error $\mathbf{u}_t$ are generated as follows:

- $\{\mathbf{b}_i\}_{i=1}^p$ are i.i.d. from $N_K(\mathbf{0}, 5\mathbf{I}_K)$.
- $\{\mathbf{f}_t\}_{t=1}^T$ are i.i.d. from $N_K(\mathbf{0}, \mathbf{I}_K)$.
- $\{\mathbf{u}_t\}_{t=1}^T$ are i.i.d. from $N_p(\mathbf{0}, 50\mathbf{I}_p)$.

The observations $\{\mathbf{y}_t\}_{t=1}^T$ are generated from (1) using $\mathbf{b}_i$, $\mathbf{f}_t$ and $\mathbf{u}_t$ from the above. Tables 1-4 report the estimation errors of the factors, the loading matrices and the covariance-of-interest $\mathbf{\Sigma}_S$ in terms of different measurements.

We see from Tables 1 and 2 that when $s = 50$ and $p = 1000, 2000$, Method 1 performs much worse than Method 2, for both $T = 200$ and $T = 400$. However, when s increases to 800 with p being the same, Tables 3 and 4 show that the improvement of Method 2 over Method 1 is less profound. This is expected as the set of interest already contains sufficiently rich information to produce an accurate estimator for realized factors. In general, we note that Method 2 is the most advantageous in the settings where s is much smaller than p. In addition, from Tables 1-4, we can tell that Method 2 comes closer to the Oracle method as p grows. In practice, we also observe that the WPC factor estimator performs better than the unweighted PC estimator when $\mathbf{u}_t$ is heteroscedastic. Due to the space limit, we choose not to present the simulation results in this model.

For further comparison with the divide-and-conquer method, we vary T from 50 to 500 and set (s, p, M) as $s = \lfloor T^{0.6} \rfloor$, $p = \lfloor T^{1.4} \rfloor$ and $M = \lfloor T^{0.2} \rfloor$. Figure 1 shows the estimation errors of the four methods together with the corresponding computational time. Again, when p is large, Method 2 performs as well as the Oracle Method, both of which greatly outperform Method 1. However, its computation becomes much slower in this case. In contrast, the divide-and-conquer method is much faster, while maintaining comparable performance as Method 2. In the extreme case that p is around 6000 ($T = 500$), the divide-and-conquer method can boost the speed by 9 fold for Method 2.

(s,p)	$(50, 1000)$			$(50, 2000)$		
Method	M1	M2	ORA	M1	M2	ORA
$\|\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|\|_{\mathbf{\Sigma}_S}$	0.271(0.014)	0.205(0.013)	0.204(0.013)	0.270(0.014)	0.201(0.013)	0.200(0.013)
$\|\|\widehat{\mathbf{\Sigma}}_S^{-1} - \mathbf{\Sigma}_S^{-1}\|\|$	0.016(0.003)	0.009(0.002)	0.009(0.002)	0.017(0.003)	0.009(0.002)	0.009(0.002)
$\|\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|\|_{\max}$	18.828(3.072)	17.460(3.237)	17.457(3.261)	18.076(2.697)	16.631(2.949)	16.623(2.950)
$\max_{t\le T} \|\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|\|$	1.811(0.195)	0.445(0.046)	NA	1.870(0.236)	0.331(0.025)	NA
$\max_{i\le s} \|\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|$	8.064(0.694)	4.100(0.330)	3.858(0.274)	8.150(0.682)	3.932(0.292)	3.805(0.297)
$\max_{i\le s, t\le T} \|\|\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\|\|$	11.375(1.262)	5.519(0.813)	5.268(0.843)	11.466(1.353)	5.253(0.776)	5.113(0.739)

Table 1: Comparison of three methods when s is much smaller than p (T = 200). M1, M2 and ORA stand for Method 1, 2 and Oracle method, respectively.

(s,p)	$(50,1000)$			$(50,2000)$		
Method	M1	M2	ORA	M1	M2	ORA
$\|\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|\|_{\boldsymbol{\Sigma}_S}$	0.186(0.009)	0.132(0.007)	0.131(0.007)	0.186(0.009)	0.131(0.008)	0.130(0.008)
$\|\|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\|\|$	0.011(0.002)	0.004(0.001)	0.004(0.001)	0.011(0.002)	0.004(0.001)	0.004(0.001)
$\|\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|\|_{\max}$	14.054(1.945)	11.922(2.245)	11.891(2.262)	14.180(2.154)	11.901(2.603)	11.900(2.604)
$\max_{t\leq T}\|\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|\|$	1.839(0.193)	0.417(0.036)	NA	1.843(0.198)	0.305(0.026)	NA
$\max_{i\leq s}\|\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|$	6.960(0.584)	2.830(0.200)	2.692(0.198)	7.024(0.605)	2.761(0.188)	2.692(0.194)
$\max_{i\leq s,t\leq T}\|\|\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\|\|$	11.871(1.540)	4.138(0.510)	3.824(0.501)	11.457(1.569)	4.088(0.516)	3.889(0.542)

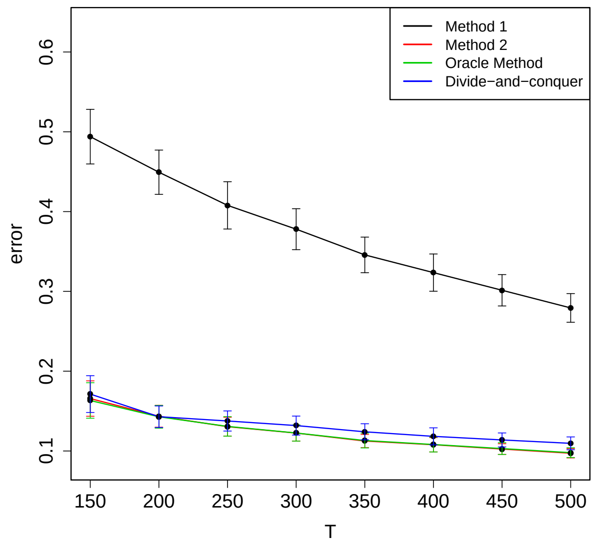
Table 2: Comparison of three methods when s is much smaller than p (T = 400). M1, M2 and ORA stand for Method 1, 2 and Oracle method, respectively.

(s,p)	$(800,1000)$			$(800,2000)$		
Method	M1	M2	ORA	M1	M2	ORA
$\|\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|\|_{\boldsymbol{\Sigma}_S}$	0.440(0.006)	0.439(0.006)	0.435(0.006)	0.439(0.006)	0.436(0.006)	0.435(0.006)
$\|\|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\|\|$	0.062(0.009)	0.062(0.009)	0.062(0.009)	0.061(0.009)	0.061(0.009)	0.062(0.012)
$\|\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|\|_{\max}$	24.565(2.626)	24.562(2.609)	24.567(2.599)	24.511(2.883)	24.543(2.847)	24.536(2.851)
$\max_{t\leq T}\|\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|\|$	0.488(0.047)	0.447(0.040)	NA	0.478(0.049)	0.337(0.038)	NA
$\max_{i\leq s}\|\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|$	15.550(0.488)	15.370(0.462)	14.418(0.271)	15.595(0.551)	15.041(0.357)	14.398(0.243)
$\max_{i\leq s,t\leq T}\|\|\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\|\|$	6.745(0.611)	6.680(0.635)	6.405(0.630)	6.904(0.734)	6.697(0.763)	6.588(0.737)

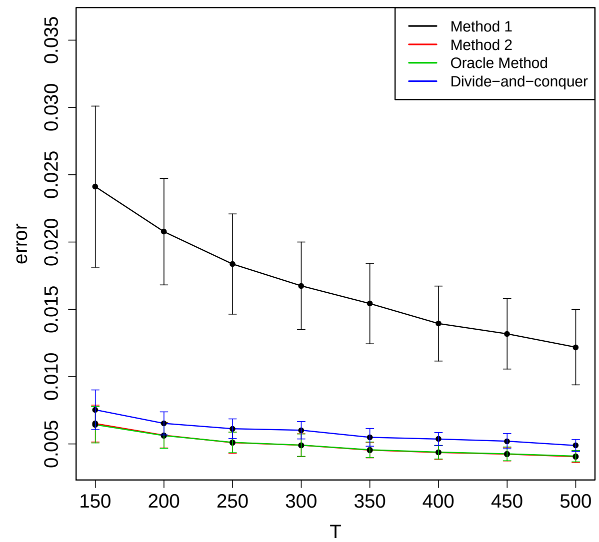
Table 3: Comparison of three methods when s is comparative to p (T = 200). M1, M2 and ORA stand for Method 1, 2 and Oracle method, respectively.

(s,p)	$(800,1000)$			$(800,2000)$		
Method	M1	M2	ORA	M1	M2	ORA
$\|\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|\|_{\boldsymbol{\Sigma}_S}$	0.193(0.004)	0.192(0.004)	0.189(0.004)	0.192(0.004)	0.190(0.004)	0.188(0.004)
$\|\|\widehat{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\|\|$	0.008(0.001)	0.008(0.001)	0.008(0.001)	0.008(0.001)	0.008(0.001)	0.008(0.001)
$\|\|\widehat{\boldsymbol{\Sigma}}_S - \boldsymbol{\Sigma}_S\|\|_{\max}$	17.062(2.603)	17.051(2.612)	17.041(2.621)	16.919(2.182)	16.891(2.206)	16.888(2.209)
$\max_{t\leq T}\|\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\|\|$	0.467(0.038)	0.423(0.036)	NA	0.466(0.038)	0.304(0.026)	NA
$\max_{i\leq s}\|\|\widehat{\mathbf{b}}_i - \mathbf{H}\mathbf{b}_i\|\|$	11.009(0.298)	10.850(0.302)	10.225(0.205)	10.934(0.274)	10.530(0.213)	10.189(0.172)
$\max_{i\leq s,t\leq T}\|\|\widehat{\mathbf{b}}_i'\widehat{\mathbf{f}}_t - \mathbf{b}_i'\mathbf{f}_t\|\|$	5.367(0.577)	5.276(0.560)	4.880(0.528)	5.293(0.411)	5.024(0.461)	4.894(0.420)

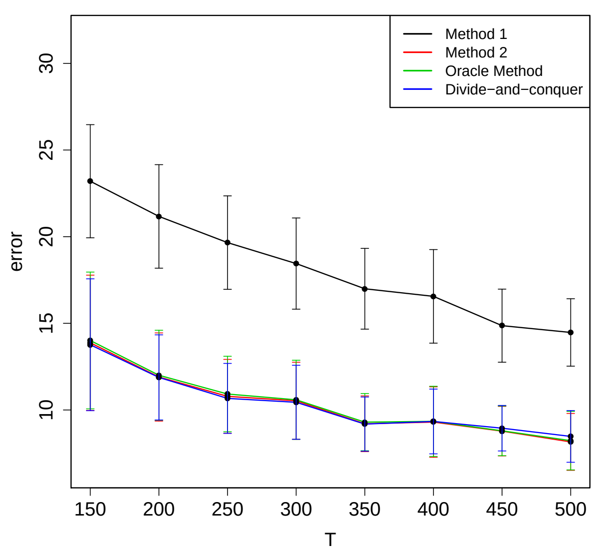
Table 4: Comparison of three methods when s is comparative to p (T = 400). M1, M2 and ORA stand for Method 1, 2 and Oracle method, respectively.



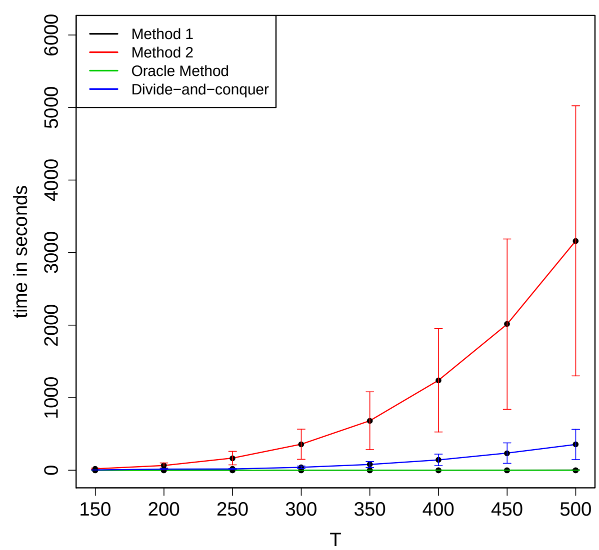
(a) $\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S}$



(b) $\|\widehat{\mathbf{\Sigma}}_S^{-1} - \mathbf{\Sigma}_S^{-1}\|$



(c) $\|\widehat{\mathbf{\Sigma}}_S - \mathbf{\Sigma}_S\|_{\max}$



(d) Computational time

Figure 1: Estimation error by four methods and their computational time: the dotted lines represent the means over 100 simulations and the segments represent the corresponding standard deviations.

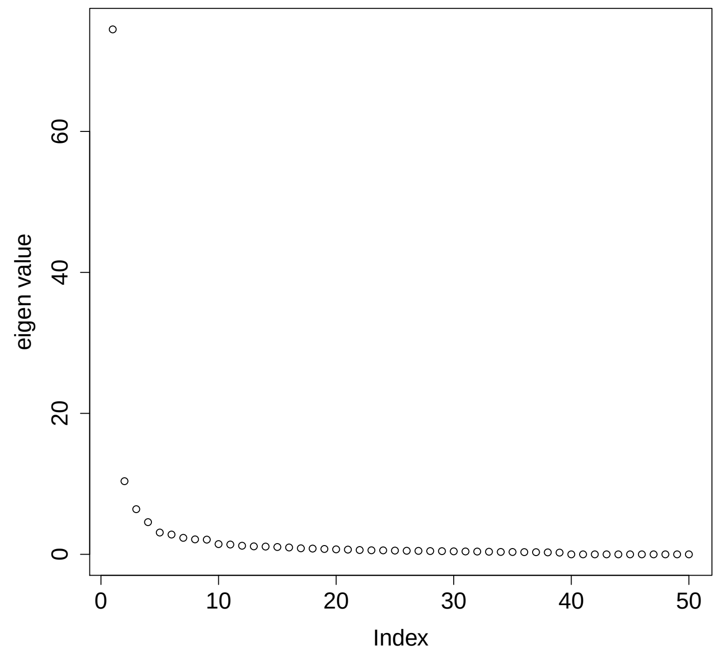


Figure 2: Eigen-values of the sample covariance matrix for GSE22255

7 Real Data Example

We use a real data example to illustrate how different utilization of available variables can affect the inference of the variables of interest. Krug et al. (2012) carried out a gene profiling study among 40 Portuguese and Spanish adults to identify key genetic risk factors for ischemic stroke. Among them, 20 subjects were patients having ischemic stroke and the others were controls. Their gene profiles were obtained using the GeneChip Human Genome U133 Plus 2.0 microarray. The data was available at Gene Expression Omnibus with access name "GSE22255".

To judge how effectively the gene expression can distinguish ischemic stroke and controls, we applied the Linear Discriminant Analysis (LDA) to this dataset. We randomly chose 10 subjects as the test set and the rest as the training set. We repeated the random splitting for 100 runs. In each run, we selected the set of expressed differentially (DE) genes with a threshold of over 1.2-fold change and a Q-value ≤ 0.05, which is a commonly used quantity to define DE genes (Storey, 2002). A LDA rule was then learned from the

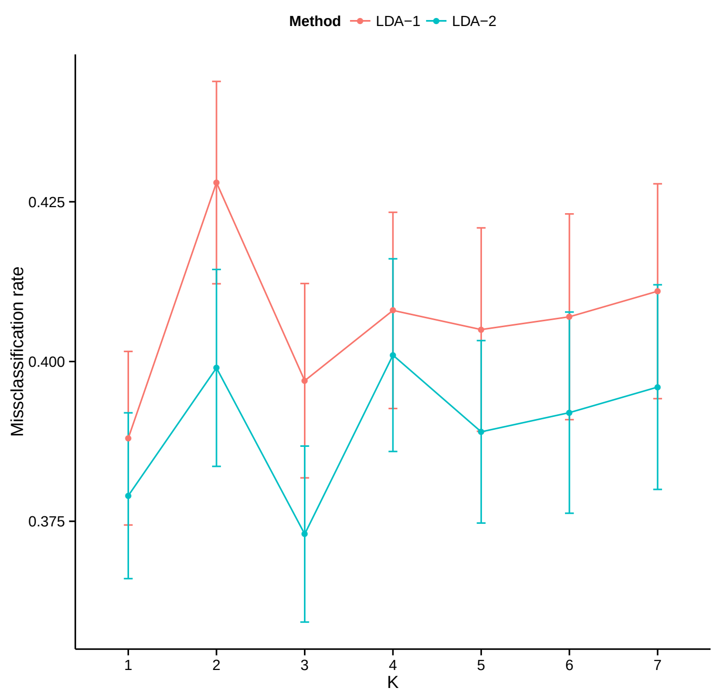


Figure 3: Misclassification rates of LDA-1 and LDA-2 over 100 random splits: the dotted lines represent the means over 100 splits and the segments represent the corresponding standard deviations.

training set using the selected genes and further applied to the test set for classifying cases and controls. The LDA rule classifies a subject as a case if

$$\widehat{\boldsymbol{\delta}}'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \bar{\boldsymbol{\mu}}) \geq 0, \qquad (11)$$

where $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0 \in \mathbb{R}^s$ is the sample mean difference between the two groups (case - control), s is the number of selected genes, $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{s \times s}$ is an estimator of the true covariance matrix $\boldsymbol{\Sigma}$ of the selected genes, and $\bar{\boldsymbol{\mu}} = (\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_0)/2$. $\bar{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\Sigma}}$ are obtained from the training set and $\mathbf{x}$ is the gene expression of subjects in the test set.

As s can be larger than the sample size, the traditional LDA where $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance is no longer applicable. An alternative method to estimate $\boldsymbol{\Sigma}$ is adopting the factor model. Factor modeling is widely used in the genomics literature to model the dependencies among genes (Carvalho et al., 2012; Kustra et al., 2006). Several factors, like the natural pathway structure (Ogata et al., 2000) can be the latent factors affecting the correlation among genes. A few spiked eigenvalues of the sample covariance in Figure 2 also suggest the existence of potential latent factors in this dataset. Again, there are two ways utilizing the factor model. One way is to use Method 1, where all procedures are done based on the selected genes only. The resulting rule is referred as "LDA-1" in Figure 3. Another way is to use auxiliary data as in Method 2. More specifically, it firstly uses data from all involved genes and subjects in the training set to estimate the latent factors. These estimated factors are then applied to the set of selected genes, where their loadings and idiosyncratic matrix estimators are obtained. Combing them together produces the covariance matrix estimator, which is still an $s \times s$ matrix. The resulting rule is referred as "LDA-2" in Figure 3. Recall that the only difference between the two rules is that they use different covariance estimators.

Figure 3 plots the average misclassification rates on the test set against the number of factors for the 100 random splits. It is clearly seen that LDA-2 gives better misclassification rates than LDA-1, which is solely due to a different estimation of the covariance matrix. The results lend further support to our claim that using more data is beneficial.

Appendix

A.1 Additional Regularity Conditions

(iv) $\{\mathbf{u}_t, \mathbf{f}_t\}_{t\geq 1}$ are i.i.d. sub-Gaussian random variables over t.

(v) There exist constants c_1 and c_2 that $0 < c_1 \leq \lambda_{\min}(\mathbf{\Sigma}_u) \leq \lambda_{\max}(\mathbf{\Sigma}_u) \leq c_2 < \infty$, $\|\mathbf{\Sigma}_u\|_1 < c_2$ and $\min_{i\leq p, j\leq p} \mathrm{Var}(u_{it}u_{jt}) > c_1$;

(vi) There exists an $M > 0$ such that $||\mathbf{B}||_{\max} < M$;

(vii) There exists an $M > 0$ such that for any $s \leq T$ and $t \leq T$, $\mathrm{E}|p^{-1/2}(\mathbf{u}_s'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t - \mathrm{E}\mathbf{u}_s'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t)|^4 < M$ and $\mathrm{E}||p^{-1/2}\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t||^4 < M$;

(viii) For each $t \leq T$, $\mathrm{E}\|(pT)^{-1/2}\sum_{s=1}^T \mathbf{f}_s(\mathbf{u}_s'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t - \mathrm{E}(\mathbf{u}_s'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t))\|^2 = O(1)$;

(ix) For each $i \leq p$, $\mathrm{E}\|(pT)^{-1/2}\sum_{t=1}^T\sum_{j=1}^p \mathbf{d}_j(u_{jt}u_{it} - \mathrm{E}u_{jt}u_{it})\| = O(1)$, where $\mathbf{d}_j$ is the jth column of $\mathbf{B}'\mathbf{\Sigma}_u^{-1}$;

(x) For each $i \leq K$, $\mathrm{E}\|(pT)^{-1/2}\sum_{t=1}^T\sum_{j=1}^N \mathbf{d}_j u_{jt} f_{it}\| = O(1)$.

Condition (iv) is a standard assumption in order to establish the exponential type of concentration inequality for the elements in $\mathbf{u}_t$ and $\mathbf{f}_t$. Condition (v) requires $\mathbf{\Sigma}_u$ to be well-conditioned. In particular, we need a lower bound on the eigen-values of $\mathbf{\Sigma}_u$. This assumption guarantees that $\tilde{\mathbf{\Sigma}}_u$ is asymptotically non-singular so that $\tilde{\mathbf{\Sigma}}_u^{-1}$ will not perform badly in the weighted least-squares problem described in (6). These conditions were also assumed in Fan et al. (2013). Conditions (vii)-(x) are some moment conditions needed to establish the central limit theorem for the WPC estimator $\widehat{\mathbf{f}}_t$. They are standard in the factor model literature, e.g. Stock and Watson (2002) and Bai (2003).

A.2 Proofs of Results in Sections 2 and 4

Proof of Proposition 1. Let $\mathbf{g}_1 = \nabla_{\boldsymbol{\theta}_S} \log h(\mathbf{y}_S - \boldsymbol{\theta}_S, \mathbf{y}_{S^c} - \boldsymbol{\theta}_{S^c})$ and $\mathbf{g}_2 = \nabla_{\boldsymbol{\theta}_S} \log h_S(\mathbf{y}_S - \boldsymbol{\theta}_S)$, where h_S is the marginal density of $\mathbf{y}_S$. Firstly, we show that $\mathbf{g}_2 = \mathrm{E}(\mathbf{g}_1|\mathbf{y}_S)$. In fact, for any bounded function $\varphi(\mathbf{y}_S)$, by Fubini Theorem and condition (3),

$$\begin{aligned}\mathrm{E}(\mathbf{g}_1\varphi(\mathbf{y}_S)) &= -\iint (\nabla_{\mathbf{y}_S} \log h(\mathbf{y}_S - \boldsymbol{\theta}_S, \mathbf{y}_{S^c} - \boldsymbol{\theta}_{S^c}))h(\mathbf{y}_S - \boldsymbol{\theta}_S, \mathbf{y}_{S^c} - \boldsymbol{\theta}_{S^c})\varphi(\mathbf{y}_S)\mathrm{d}\mathbf{y}_S\mathrm{d}\mathbf{y}_{S^c} \\ &= -\iint (\nabla_{\mathbf{y}_S} h(\mathbf{y}_S - \boldsymbol{\theta}_S, \mathbf{y}_{S^c} - \boldsymbol{\theta}_{S^c}))\varphi(\mathbf{y}_S)\mathrm{d}\mathbf{y}_S\mathrm{d}\mathbf{y}_{S^c}\end{aligned}$$

$$
\begin{aligned}
&= -\int \left(\nabla_{\mathbf{y}_S} \int h(\mathbf{y}_S - \boldsymbol{\theta}_S, \mathbf{y}_{S^c} - \boldsymbol{\theta}_{S^c}) \mathrm{d}\mathbf{y}_{S^c} \right) \varphi(\mathbf{y}_S) \mathrm{d}\mathbf{y}_S \\
&= -\int \nabla_{\mathbf{y}_S} h_S(\mathbf{y}_S - \boldsymbol{\theta}_S) \varphi(\mathbf{y}_S) \mathrm{d}\mathbf{y}_S \\
&= \int \left(\nabla_{\mathbf{y}_S} \log h_S(\mathbf{y}_S - \boldsymbol{\theta}_S) \right) h_S(\mathbf{y}_S - \boldsymbol{\theta}_S) \varphi(\mathbf{y}_S) \mathrm{d}\mathbf{y}_S \\
&= \mathrm{E}(\mathbf{g}_2 \varphi(\mathbf{y}_S)).
\end{aligned}
$$

Then, by definition, $\mathbf{g}_2 = \mathrm{E}(\mathbf{g}_1 | \mathbf{y}_S)$. Therefore,

$$
\begin{aligned}
\{I_p(\boldsymbol{\theta})\}_S &= \mathrm{E}(\mathbf{g}_1 \mathbf{g}_1') = \mathrm{E}[(\mathbf{g}_2 + \mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_2 + \mathbf{g}_1 - \mathbf{g}_2)'] \\
&= \mathrm{E}[\mathbf{g}_2 \mathbf{g}_2'] + \mathrm{E}[\mathbf{g}_2(\mathbf{g}_1 - \mathbf{g}_2)'] + \mathrm{E}[(\mathbf{g}_1 - \mathbf{g}_2)\mathbf{g}_2'] + \mathrm{E}[(\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)'] \\
&= I_S(\boldsymbol{\theta}_S) + \mathrm{E}[(\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)'] \\
&\succeq I_S(\boldsymbol{\theta}_S),
\end{aligned}
$$

where the last equality follows from $\mathrm{E}[\mathbf{g}_2(\mathbf{g}_1 - \mathbf{g}_2)'] = \mathrm{E}[\mathrm{E}[\mathbf{g}_2(\mathbf{g}_1 - \mathbf{g}_2)' | \mathbf{y}_S]] = 0$, since $\mathbf{g}_2 = \mathrm{E}(\mathbf{g}_1 | \mathbf{y}_S)$. □

Proof of Example 2. Without loss of generality, we assume $\boldsymbol{\theta} = \mathbf{0}$ so that the density of $\mathbf{y}$ is proportional to $g(\mathbf{y}'\boldsymbol{\Omega}\mathbf{y})$, where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Then,

$$
\begin{aligned}
|\nabla_{\mathbf{y}_S} h(\mathbf{y}_S, \mathbf{y}_{S^c})| &= 2\,|g'(\mathbf{y}'\boldsymbol{\Omega}\mathbf{y})(\boldsymbol{\Omega}\mathbf{y})_S| \leq 2\,|g'(\mathbf{y}'\boldsymbol{\Omega}\mathbf{y})|\,|\boldsymbol{\Omega}_S \mathbf{y}_S + \boldsymbol{\Omega}_{S,S^c} \mathbf{y}_{S^c}| \\
&\leq 2c\,|\boldsymbol{\Omega}_S \mathbf{y}_S + \boldsymbol{\Omega}_{S,S^c} \mathbf{y}_{S^c}|\, g(\mathbf{y}'\boldsymbol{\Omega}\mathbf{y}).
\end{aligned}
$$

Note that

$$
\begin{aligned}
\int \left(\int |\boldsymbol{\Omega}_S \mathbf{y}_S + \boldsymbol{\Omega}_{S,S^c} \mathbf{y}_{S^c}|\, g(\mathbf{y}'\boldsymbol{\Omega}\mathbf{y}) d\mathbf{y}_{S^c} \right) d\mathbf{y}_S &\propto \mathrm{E}\left(|\boldsymbol{\Omega}_S \mathbf{y}_S + \boldsymbol{\Omega}_{S,S^c} \mathbf{y}_{S^c}|\right) \\
&\leq \mathrm{E}\left(|\boldsymbol{\Omega}_S \mathbf{y}_S| + |\boldsymbol{\Omega}_{S,S^c} \mathbf{y}_{S^c}|\right) \\
&< \infty
\end{aligned}
$$

Therefore for a.e. any $\mathbf{y}_S$, $\int |\boldsymbol{\Omega}_S \mathbf{y}_S + \boldsymbol{\Omega}_{S,S^c} \mathbf{y}_{S^c}|\, g(\mathbf{y}'\boldsymbol{\Omega}\mathbf{y})$ is integrable. By Example 1.8 of Shao (2003), differentiation and integration are interchangeable, hence (3) holds. □

Proof of Proposition 2. For simplicity, let $\boldsymbol{\Omega} = I_p(\boldsymbol{\theta})$ and partition it as

$$
\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_S & \boldsymbol{\Omega}_{S,S^c} \\ \boldsymbol{\Omega}_{S^c,S} & \boldsymbol{\Omega}_{S^c} \end{pmatrix}.
$$

Then, the Fisher information $I(\mathbf{f})$ of $\mathbf{f}$ contained in all data is given by

$$I(\mathbf{f}) = \mathbf{B}'\mathbf{\Omega}\mathbf{B} = \mathbf{B}_S'\mathbf{\Omega}_S\mathbf{B}_S + \mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c,S}\mathbf{B}_S + \mathbf{B}_S'\mathbf{\Omega}_{S,S^c}\mathbf{B}_{S^c} + \mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c}\mathbf{B}_{S^c}. \tag{A.1}$$

If $\mathbf{\Omega}_{S,S^c} = \mathbf{0}$, we have

$$\begin{aligned} I(\mathbf{f}) &= \mathbf{B}_S'\mathbf{\Omega}_S\mathbf{B}_S + \mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c}\mathbf{B}_{S^c} = \mathbf{B}_S'\{I_p(\boldsymbol{\theta})\}_S\mathbf{B}_S + \mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c}\mathbf{B}_{S^c} \\ &\succeq \mathbf{B}_S' I_S(\boldsymbol{\theta}_S)\mathbf{B}_S + \mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c}\mathbf{B}_{S^c} \succeq \mathbf{B}_S' I_S(\boldsymbol{\theta}_S)\mathbf{B}_S = I_S(\mathbf{f}), \end{aligned}$$

where the first inequality follows from Proposition 1 and the last inequality follows from that $\mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c}\mathbf{B}_{S^c}$ is positive semi-definite. This completes the proof. □

Proof of Proposition 3. For any general $\mathbf{Q} \in \mathbb{R}^{L\times R}$, $\mathbf{B}_L \in \mathbb{R}^{L\times K}$, and $\mathbf{B}_R \in \mathbb{R}^{R\times K}$, we have

$$\mathrm{E}(\mathbf{B}_L'\mathbf{Q}\mathbf{B}_R) = \mathrm{E}\left[\sum_{l=1}^{L}\sum_{r=1}^{R} q_{l,r}\mathbf{b}_{L,l}\mathbf{b}_{R,r}'\right].$$

where $q_{l,r}$ is the (l,r)-th element of $\mathbf{Q}$, $\mathbf{b}_{L,l}'$ is the lth row of $\mathbf{B}_L$ and $\mathbf{b}_{R,r}'$ is the rth row of $\mathbf{B}_R$. Therefore,

$$\mathrm{E}(\mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c,S}\mathbf{B}_S) = \mathrm{E}\left[\sum_{l\in S^C}\sum_{r\in S}\omega_{l,r}\mathbf{b}_{S^c,l}\mathbf{b}_{S,r}'\right],$$

where $\omega_{l,r}$ is the (l,r)-th element of $\mathbf{\Omega}$. By the i.i.d assumption, for $l \in S^C$ and $r \in S$, $\mathrm{E}(\mathbf{b}_{S^c,l}\mathbf{b}_{S,r}') = \mathrm{E}(\mathbf{b}_{S^c,l})\mathrm{E}(\mathbf{b}_{S,r}') = \mathbf{0}$. Hence, $\mathrm{E}(\mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c,S}\mathbf{B}_S) = \mathbf{0}$. Similarly, it can be shown that $\mathrm{E}(\mathbf{B}_S'\mathbf{\Omega}_{S,S^c}\mathbf{B}_{S^c}) = \mathbf{0}$. By Proposition 1, $\mathbf{B}_S'\mathbf{\Omega}_S\mathbf{B}_S \succeq \mathbf{I}_S(\mathbf{f})$, which implies that $\mathrm{E}(\mathbf{B}_S'\mathbf{\Omega}_S\mathbf{B}_S) \succeq \mathrm{E}(\mathbf{I}_S(\mathbf{f}))$.

$$\mathrm{E}(\mathbf{B}_{S^c}'\mathbf{\Omega}_{S^c}\mathbf{B}_{S^c}) = \mathrm{E}\left[\sum_{l\in S^c}\sum_{r\in S^c}\omega_{l,r}\mathbf{b}_{L,l}\mathbf{b}_{R,r}'\right] = \mathrm{E}\left[\sum_{l\in S^c}\omega_{l,l}\mathbf{b}_{L,l}\mathbf{b}_{L,l}'\right] = \mathrm{tr}(\mathbf{\Omega}_{S^c})\mathrm{E}(\mathbf{b}\mathbf{b}') \succeq \mathbf{0}.$$

Using (A.1) and the above results, we have $\mathrm{E}[I(\mathbf{f})] \succeq \mathrm{E}[I_S(\mathbf{f})]$. □

Proof of Lemma 1. Since we assume all conditions hold for both s and p, we prove the result for p, i.e. $\max_{t\le T}\|\widehat{\mathbf{f}}_t^{(2)} - \mathbf{H}_2\mathbf{f}_t\| = O_P\left(T^{-1/2} + T^{1/4}/p^{-1/2}\right)$. The result for s can be proved similarly. For simplicity, we write $\widehat{\mathbf{f}}_t^{(2)}$ as $\widehat{\mathbf{f}}_t$ and $\mathbf{H}_2$ as $\mathbf{H}$.

By (A.1) of Bai and Liao (2013), $\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t$ has the following expansion,

$$\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t = \widehat{\mathbf{V}}^{-1}\left(\frac{1}{T}\sum_{i=1}^{T}\widehat{\mathbf{f}}_i\mathbf{u}_i'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{u}_t/p + \frac{1}{T}\sum_{i=1}^{T}\widehat{\mathbf{f}}_i\widehat{\eta}_{it} + \frac{1}{T}\sum_{i=1}^{T}\widehat{\mathbf{f}}_i\widehat{\theta}_{it}\right),$$

where $\widehat{\eta}_{it} = \mathbf{f}_i'\mathbf{B}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{u}_t/p$, $\widehat{\theta}_{it} = \mathbf{f}_t'\mathbf{B}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{u}_i/p$, and $\widehat{\mathbf{V}}$ is the diagonal matrix of the K largest eigenvalues of $\mathbf{Y}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{Y}/T$. Let $\eta_{it} = \mathbf{f}_i'\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t/p$ and $\theta_{it} = \mathbf{f}_t'\mathbf{B}'\mathbf{\Sigma}_u^{-1}\mathbf{u}_i/p$. Then, we have

$$\begin{aligned}\|\widehat{\mathbf{f}}_t - \mathbf{H}\mathbf{f}_t\| \leq \|\widehat{\mathbf{V}}^{-1}\| \Bigg(& \left\|\frac{1}{T}\sum_{i=1}^T \widehat{\mathbf{f}}_i\mathbf{u}_i'(\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1})\mathbf{u}_t/p\right\| + \left\|\frac{1}{T}\sum_{i=1}^T \widehat{\mathbf{f}}_i(\mathbf{u}_i'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t - \mathrm{E}\mathbf{u}_i'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t)/p\right\| \\ & + \left\|\frac{1}{T}\sum_{i=1}^T \widehat{\mathbf{f}}_i\mathrm{E}(\mathbf{u}_i'\mathbf{\Sigma}_u^{-1}\mathbf{u}_t)/p\right\| + \left\|\frac{1}{T}\sum_{i=1}^T \widehat{\mathbf{f}}_i(\widehat{\eta}_{it} - \eta_{it})\right\| + \left\|\frac{1}{T}\sum_{i=1}^T \widehat{\mathbf{f}}_i\eta_{it}\right\| \\ & + \left\|\frac{1}{T}\sum_{i=1}^T \widehat{\mathbf{f}}_i(\widehat{\theta}_{it} - \theta_{it})\right\| + \left\|\frac{1}{T}\sum_{i=1}^T \widehat{\mathbf{f}}_i\theta_{it}\right\|\Bigg). \end{aligned} \tag{A.2}$$

Denote the jth summand inside the parenthesis as G_{jt}.

By Lemma A.2 of Bai and Liao (2013), $\|\widehat{\mathbf{V}}^{-1}\| = O_P(1)$. By Lemma A.6(iv) of Bai and Liao (2013),

$$\max_{t\leq T} G_{1t} = O_P\left(\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| \left\{\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| + 1/\sqrt{p} + \sqrt{(\log p)/T}\right\}\right).$$

By Proposition 4.1 of Bai and Liao (2013),

$$\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| = o_P\left(\min\left\{T^{-1/4}, p^{-1/4}, \sqrt{T/(p\log p)}\right\}\right), \tag{A.3}$$

therefore, $\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| \left(\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| + 1/\sqrt{p} + \sqrt{(\log p)/T}\right) = o(T^{-1/2} + p^{-1/2})$. Hence,

$$\max_{t\leq T} G_{1t} = o_P\left(T^{-1/2} + p^{-1/2}\right).$$

By Lemma A.8(ii) of Bai and Liao (2013), $\max_{t\leq T} G_{2t} = O_P\left(T^{1/4}p^{-1/2}\right)$. By Lemma A.10(i) of Bai and Liao (2013), $\max_{t\leq T} G_{3t} = O_P\left(T^{-1/2}\right)$. By Lemma A.6(vi) of Bai and Liao (2013),

$$\max_{t\leq T} G_{4t} = O_P\left(\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| \left\{\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| + 1/\sqrt{p} + 1/\sqrt{T}\right\}\right) + o_P(1/\sqrt{p}) = o_P(1/\sqrt{p}).$$

By Lemma A.8(iii) of Bai and Liao (2013), $\max_{t\leq T} G_{5t} = O_P\left(T^{1/4}p^{-1/2}\right)$. By Lemma A.6(v) of Bai and Liao (2013) and (A.3),

$$\max_{t\leq T} G_{6t} = O_P\left(\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| \left\{\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| + 1/\sqrt{p} + \sqrt{(\log p)/T}\right\}\right) = o_P(1/\sqrt{p}).$$

By Lemma A.6(iii) of Bai and Liao (2013) and (A.3),

$$\max_{t\leq T} G_{7t} = O_P\left(\|\tilde{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\|/\sqrt{p} + 1/p + 1/\sqrt{pT}\right) = o_P(1/\sqrt{p}).$$

Then, by (A.2), we have

$$\max_{t\le T}\|\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t\|=O_P\left(\frac{1}{\sqrt{T}}+\frac{T^{1/4}}{\sqrt{p}}\right).$$

□

Proof of Lemma 2. For Method 1, we have the following decomposition

$$\widehat{\mathbf{b}}_i^{(1)}-\mathbf{H}_1\mathbf{b}_i=\underbrace{\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_1\mathbf{f}_t u_{it}}_{I_1}+\underbrace{\frac{1}{T}\sum_{t=1}^{T}y_{it}(\widehat{\mathbf{f}}_t^{(1)}-\mathbf{H}_1\mathbf{f}_t)}_{I_2}+\underbrace{\mathbf{H}_1(\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t'-\mathbf{I}_K)\mathbf{b}_i}_{I_3},$$

where $\mathbf{b}_i$ is the true factor loading of the ith subject as defined in (1).

For I_1, we have

$$\max_{i\le s}\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_1\mathbf{f}_t u_{it}\right\|\le\|\mathbf{H}_1\|\max_{i\le s}\sqrt{\sum_{k=1}^{K}\Big(\frac{1}{T}\sum_{t=1}^{T}f_{kt}u_{it}\Big)^2}.$$

It follows from Lemma C.3(iii) of Fan et al. (2013) that, $\max_{i\le s}\sqrt{\sum_{k=1}^{K}(\frac{1}{T}\sum_{t=1}^{T}f_{kt}u_{it})^2}$ $=O_P\left(\sqrt{(\log s)/T}\right)$. From Lemma A.2, $\|\mathbf{H}_1\|=O_P(1)$, therefore $I_1=O_P\left(\sqrt{(\log s)/T}\right)$.

As for I_2, by conditions (v) and (vi),

$$\max_{i\le s}\mathrm{E}y_{it}^2=\max_{i\le s}\{\mathrm{E}(\mathbf{b}_i'\mathbf{f}_t)^2+\mathrm{E}u_{it}^2\}\le\max_{i\le s}\|\mathbf{b}_i\|^2+\max_{i\le s}\mathrm{Var}(u_{it})=O(1).$$

By condition (iv), y_{it}^2 is sub-exponential, therefore by the union bound and sub-exponential tail bound, $\max_{i\le s}\left|\frac{1}{T}\sum_{t=1}^{T}y_{it}^2-\mathrm{E}y_{it}^2\right|=O_P\left(\sqrt{(\log s)/T}\right)$. Then,

$$\max_{i\le s}\frac{1}{T}\sum_{t=1}^{T}y_{it}^2\le\max_{i\le s}\left|\frac{1}{T}\sum_{t=1}^{T}y_{it}^2-\mathrm{E}y_{it}^2\right|+\max_{i\le s}\mathrm{E}y_{it}^2=O_P(1).\tag{A.4}$$

By Cauchy-Schwartz inequality,

$$\begin{aligned}\max_{i\le s}\left\|\frac{1}{T}\sum_{t=1}^{T}y_{it}(\widehat{\mathbf{f}}_t^{(1)}-\mathbf{H}_1\mathbf{f}_t)\right\|&\le\max_{i\le s}\left(\frac{1}{T}\sum_{t=1}^{T}y_{it}^2\cdot\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t^{(1)}-\mathbf{H}_1\mathbf{f}_t\|^2\right)^{1/2}\\&=O_P\left(\left(\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t^{(1)}-\mathbf{H}_1\mathbf{f}_t\|^2\right)^{1/2}\right)\\&=O_P\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{s}}\right),\end{aligned}$$

where the last equality follows from Lemma A.5. So, $I_2 = O_P\left(1/\sqrt{T} + 1/\sqrt{s}\right)$.

Finally, it follows from Lemma C.3(i) of Fan et al. (2013) that $\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\| = O_P\left(T^{-1/2}\right)$. This together with $\|\mathbf{H}_1\| = O_P(1)$ and condition (vi) show that $I_3 = O_P\left(T^{-1/2}\right)$. Hence,

$$\max_{i\leq s}\|\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}_1\mathbf{b}_i\| = O_P\left(\frac{1}{\sqrt{s}} + \sqrt{\frac{\log s}{T}}\right).$$

Using the same arguments and the results of $\widehat{\mathbf{f}}_t^{(2)}$ in Lemma 1, we can show that

$$\max_{i\leq s}\|\widehat{\mathbf{b}}_i^{(2)} - \mathbf{H}_2\mathbf{b}_i\| = O_P\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log s}{T}}\right).$$

When the common factor $\mathbf{f}_t$ is known, for the oracle estimator of the loading matrix, we have

$$\begin{aligned}\max_{i\leq s}\|\widehat{\mathbf{b}}_i^{o} - \mathbf{b}_i\| &\leq \max_{i\leq s}\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t u_{it}\right\| + \left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t\mathbf{f}_t' - \mathbf{I}_K\right\|\max_{i\leq s}\|\mathbf{b}_i\| \\ &= O_P\left(\sqrt{\frac{\log s}{T}} + \frac{1}{\sqrt{T}}\right) \\ &= O_P\left(\sqrt{\frac{\log s}{T}}\right).\end{aligned}$$

□

Proof of Lemma 3. By Theorem A.1 of Fan et al. (2013) (cited as Lemma A.7 in Appendix), it suffices to show

$$\max_{i\leq s}\frac{1}{T}\sum_{t=1}^{T}(u_{it} - \widehat{u}_{it}^{(1)})^2 = O_P\left(\frac{1}{s} + \frac{\log s}{T}\right) \quad \text{and} \quad \max_{i,t}|u_{it} - \widehat{u}_{it}^{(1)}| = o_P(1).$$

For Method 1, we have

$$u_{it} - \widehat{u}_{it}^{(1)} = \mathbf{b}_i'\mathbf{H}_1'(\widehat{\mathbf{f}}_t^{(1)} - \mathbf{H}_1\mathbf{f}_t) + \{(\widehat{\mathbf{b}}_i^{(1)})' - \mathbf{b}_i'\mathbf{H}_1\}\widehat{\mathbf{f}}_t^{(1)} + \mathbf{b}_i'(\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K)\mathbf{f}_t$$

Using $(a+b+c)^2 \leq 4a^2 + 4b^2 + 4c^2$, we have

$$\begin{aligned}\max_{i\leq s}\frac{1}{T}\sum_{t=1}^{T}(u_{it} - \widehat{u}_{it}^{(1)})^2 \leq{}& 4\max_{i\leq s}\|\mathbf{H}_1\mathbf{b}_i\|^2\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t^{(1)} - \mathbf{H}_1\mathbf{f}_t\|^2 \\ &+ 4\max_{i\leq s}\|\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}_1\mathbf{b}_i\|^2\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t^{(1)}\|^2\end{aligned}$$

$$+ 4\|\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K\|_F^2 \max_{i\le s}\|\mathbf{b}_i\|^2 \frac{1}{T}\sum_{t=1}^T \|\mathbf{f}_t\|^2.$$

Since, $\max_i\|\mathbf{H}_1\mathbf{b}_i\| \le \|\mathbf{H}_1\| \max_i\|\mathbf{b}_i\| = O_P(1)$, $\frac{1}{T}\sum_{t=1}^T\|\widehat{\mathbf{f}}_t^{(1)}\|^2 = O_P(1)$, and $\frac{1}{T}\sum_{t=1}^T\|\mathbf{f}_t\|^2 = O_P(1)$, it follows from Lemma 1, 2, A.3 and A.5 that

$$\max_{i\le s}\frac{1}{T}\sum_{t=1}^T (u_{it} - \widehat{u}_{it}^{(1)})^2 = O_P\left(\frac{1}{s} + \frac{\log s}{T}\right). \tag{A.5}$$

On the other hand, by Lemma A.1,

$$\max_{i,t}|u_{it} - \widehat{u}_{it}^{(1)}| = \max_{i,t}|(\widehat{\mathbf{b}}_i^{(1)})'\widehat{\mathbf{f}}_i^{(1)} - \mathbf{b}_i'\mathbf{f}_t| = O_P\left((\log T)^{1/2}\sqrt{\frac{\log s}{T}} + \frac{T^{1/4}}{\sqrt{s}}\right) = o(1).$$

Then, the result follows from Theorem A.1 of Fan et al. (2013).

In analogous, a similar result can be proved for Method 2. For the oracle estimator, $\widehat{u}_{it}^o = y_{it} - (\widehat{\mathbf{b}}_i^o)'\mathbf{f}_t$. Therefore,

$$\max_{i\le s}\frac{1}{T}\sum_{t=1}^T (u_{it} - \widehat{u}_{it}^o)^2 \le \max_{i\le s}\|\widehat{\mathbf{b}}_i^o - \mathbf{b}_i\|^2 \frac{1}{T}\sum_{t=1}^T\|\mathbf{f}_t\|^2 = O_P\left(\max_{i\le s}\|\widehat{\mathbf{b}}_i^o - \mathbf{b}_i\|^2\right) = O_P\left(\frac{\log s}{T}\right).$$

$$\max_{i,t}|u_{it} - \widehat{u}_{it}^o| = \max_{i,t}|(\widehat{\mathbf{b}}_i^o)'\mathbf{f}_t - \mathbf{b}_i'\mathbf{f}_t| = O_P\left((\log T)^{1/2}\sqrt{\frac{\log s}{T}}\right) = o_P(1).$$

It then follows from Theorem A.1 of Fan et al. (2013) that

$$\|\widehat{\mathbf{\Sigma}}_{u,S}^o - \mathbf{\Sigma}_{u,S}\| = O_P\left(m_s\sqrt{\frac{\log s}{T}}\right) = \|(\widehat{\mathbf{\Sigma}}_{u,S}^o)^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\|.$$

□

Proof of Theorem 1. (1) For Method 1, $\widehat{\mathbf{\Sigma}}_S^{(1)} = \widehat{\mathbf{B}}_1\widehat{\mathbf{B}}_1' + \widehat{\mathbf{\Sigma}}_{u,S}^{(1)}$. Therefore,

$$\begin{aligned}\|\widehat{\mathbf{\Sigma}}_S^{(1)} - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S}^2 &\le 2\left(\|\widehat{\mathbf{B}}_1\widehat{\mathbf{B}}_1' - \mathbf{B}_S\mathbf{B}_S'\|_{\mathbf{\Sigma}_S}^2 + \|\widehat{\mathbf{\Sigma}}_{u,S}^{(1)} - \mathbf{\Sigma}_{u,S}\|_{\mathbf{\Sigma}_S}^2\right)\\ &\le 2\big(\|\mathbf{B}_S(\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K)\mathbf{B}_S'\|_{\mathbf{\Sigma}_S}^2 + 2\|\mathbf{B}_S\mathbf{H}_1'\mathbf{C}_1'\|_{\mathbf{\Sigma}_S}^2 + \|\mathbf{C}_1\mathbf{C}_1'\|_{\mathbf{\Sigma}_S}^2\\ &\quad + \|\widehat{\mathbf{\Sigma}}_{u,S}^{(1)} - \mathbf{\Sigma}_{u,S}\|_{\mathbf{\Sigma}_S}^2\big),\end{aligned}$$

where $\mathbf{C}_1 = \widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1'$. Then, it follows from Lemmas A.4 that

$$\|\widehat{\mathbf{\Sigma}}_S^{(1)} - \mathbf{\Sigma}_S\|_{\mathbf{\Sigma}_S}^2 = O_P\left(\frac{1}{sT} + \frac{1}{s^2} + w_1^2 + sw_1^4 + m_s^2w_1^2\right) = O_P\left(sw_1^4 + m_s^2w_1^2\right).$$

Similarly, $\|\widehat{\boldsymbol{\Sigma}}_S^{(2)} - \boldsymbol{\Sigma}_S\|_{\boldsymbol{\Sigma}_S}^2 = O_P\left(sw_2^4 + m_s^2 w_2^2\right)$.

In the oracle case, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}_S^o - \boldsymbol{\Sigma}_S\|_{\boldsymbol{\Sigma}_S}^2 &\leq 2\left(\|\widehat{\mathbf{B}}_o\widehat{\mathbf{B}}_o' - \mathbf{B}_S\mathbf{B}_S'\|_{\boldsymbol{\Sigma}_S}^2 + \|\widehat{\boldsymbol{\Sigma}}_{u,S}^o - \boldsymbol{\Sigma}_{u,S}\|_{\boldsymbol{\Sigma}_S}^2\right) \\
&\leq 2\Big(\underbrace{\|(\widehat{\mathbf{B}}_o - \mathbf{B}_S)(\widehat{\mathbf{B}}_o - \mathbf{B}_S)'\|_{\boldsymbol{\Sigma}_S}^2}_{I_1} + 2\underbrace{\|(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\mathbf{B}_S'\|_{\boldsymbol{\Sigma}_S}^2}_{I_2} + \underbrace{\|\widehat{\boldsymbol{\Sigma}}_{u,S}^o - \boldsymbol{\Sigma}_{u,S}\|_{\boldsymbol{\Sigma}_S}^2}_{I_3}\Big).
\end{aligned}$$

Since all eigenvalues of $\boldsymbol{\Sigma}_S$ are bounded away from zero, for any matrix $\mathbf{A} \in \mathbb{R}^{s\times s}$, $\|\mathbf{A}\|_{\boldsymbol{\Sigma}_S}^2 = s^{-1}\|\boldsymbol{\Sigma}^{-1/2}\mathbf{A}\boldsymbol{\Sigma}^{-1/2}\|_F^2 = O_P\left(s^{-1}\|\mathbf{A}\|_F^2\right)$. Then, by Lemma 2, we have

$$I_1 = O_P\left(s^{-1}\|\widehat{\mathbf{B}}_o - \mathbf{B}_S\|_F^4\right) = O_P\left(sw_o^4\right),$$

where the last equality follows that $\|\widehat{\mathbf{B}}_o - \mathbf{B}_S\|_F^2 \leq s(\max_{i\leq s}\|\widehat{\mathbf{b}}_i^o - \mathbf{b}_i\|)^2 = O_P\left(sw_o^2\right)$. For I_2, we have

$$\begin{aligned}
I_2 &= s^{-1}\mathrm{tr}((\widehat{\mathbf{B}}_o - \mathbf{B}_S)'\boldsymbol{\Sigma}_S^{-1}(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\mathbf{B}_S'\boldsymbol{\Sigma}_S^{-1}\mathbf{B}_S) \\
&\leq s^{-1}\|\boldsymbol{\Sigma}_S^{-1}\|\|\widehat{\mathbf{B}}_o - \mathbf{B}_S\|_F^2\|\mathbf{B}_S'\boldsymbol{\Sigma}_S^{-1}\mathbf{B}_S\| \\
&= O_P\left(w_o^2\right).
\end{aligned}$$

For I_3, Lemma 3 implies that

$$I_3 = O_P\left(s^{-1}\|\widehat{\boldsymbol{\Sigma}}_{u,S}^o - \boldsymbol{\Sigma}_{u,S}\|_F^2\right) = O_P\left(\|\widehat{\boldsymbol{\Sigma}}_{u,S}^o - \boldsymbol{\Sigma}_{u,S}\|^2\right) = O_P\left(m_s^2 w_o^2\right).$$

Therefore, $\|\widehat{\boldsymbol{\Sigma}}_{u,S}^o - \boldsymbol{\Sigma}_{u,S}\|_{\boldsymbol{\Sigma}_S}^2 = O_P\left(sw_o^4 + m_s^2 w_o^2\right)$.

(2) For Method 1,

$$\|\widehat{\boldsymbol{\Sigma}}_S^{(1)} - \boldsymbol{\Sigma}_S\|_{\max} \leq \underbrace{\|\widehat{\mathbf{B}}_1\widehat{\mathbf{B}}_1' - \mathbf{B}_S\mathbf{B}_S'\|_{\max}}_{I_1} + \underbrace{\|\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)} - \boldsymbol{\Sigma}_{u,S}\|_{\max}}_{I_2}.$$

For I_1, we have

$$\begin{aligned}
I_1 &= \max_{ij}|(\widehat{\mathbf{b}}_i^{(1)})'\widehat{\mathbf{b}}_j^{(1)} - \mathbf{b}_i'\mathbf{b}_j| \\
&\leq \max_{ij}\left(|(\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}_1\mathbf{b}_i)'(\widehat{\mathbf{b}}_j^{(1)} - \mathbf{H}_1\mathbf{b}_j)| + 2|\mathbf{b}_i'\mathbf{H}_1'(\widehat{\mathbf{b}}_j^{(1)} - \mathbf{H}_1\mathbf{b}_j)| + |\mathbf{b}_i'(\mathbf{H}_1\mathbf{H}_1' - \mathbf{I}_K)\mathbf{b}_j|\right) \\
&\leq \big(\max_i\|\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}_1\mathbf{b}_i\|\big)^2 + 2\max_{ij}\|\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}_1\mathbf{b}_i\|\|\mathbf{H}_1\mathbf{b}_j\| + \|\mathbf{H}_1\mathbf{H}_1' - \mathbf{I}_K\|\big(\max_i\|\mathbf{b}_i\|\big)^2 \\
&= O_P\left(w_1\right),
\end{aligned}$$

where the last identity follows from Lemmas 2 and A.3.

For I_2, let $\sigma_{u,ij}$ be the (i,j)-th entry of $\mathbf{\Sigma}_{u,S}$ and $\widehat{\sigma}_{u,ij} = \frac{1}{T}\sum_{t=1}^{T}\widehat{u}_{it}\widehat{u}_{jt}$, where $\widehat{u}_{it}$ are the estimator of u_{it} from Method 1 as described in Section 4. Then,

$$\begin{aligned}
&\max_{ij}|\widehat{\sigma}_{u,ij}-\sigma_{u,ij}| \\
=& \max_{ij}\left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{it}\widehat{u}_{jt}-u_{it}u_{jt})\right| + \max_{ij}\left|\frac{1}{T}\sum_{i=1}^{T}u_{it}u_{jt}-\mathrm{E}(u_{it}u_{jt})\right| \\
\le& \max_{ij}\left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{it}-u_{it})(\widehat{u}_{jt}-u_{jt})\right| + 2\max_{ij}\left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{it}-u_{it})u_{jt}\right| + \max_{ij}\left|\frac{1}{T}\sum_{i=1}^{T}u_{it}u_{jt}-\mathrm{E}(u_{it}u_{jt})\right| \\
\le& \max_{ij}\left(\frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{it}-u_{it})^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{jt}-u_{jt})^2\right)^{1/2} + 2\max_{ij}\left(\frac{1}{T}\sum_{t=1}^{T}(\widehat{u}_{it}-u_{it})^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}u_{jt}^2\right)^{1/2} \\
&+ \max_{ij}\left|\frac{1}{T}\sum_{i=1}^{T}u_{it}u_{jt}-\mathrm{E}(u_{it}u_{jt})\right| \\
=& O_P\left(w_1^2\right) + O_P\left(w_1\right) + O_P\left(\sqrt{(\log s)/T}\right),
\end{aligned}$$

where the last equality follows from (A.5), Lemma C.3 (ii) of Fan et al. (2013) and

$$\max_{j\le s}\frac{1}{T}\sum_{t=1}^{T}u_{jt}^2 = O_P(1)$$

as similarly shown in (A.4). Hence, $\max_{ij}|\widehat{\sigma}_{u,ij}-\sigma_{u,ij}| = O_P(w_1)$. After the thresholding,

$$\begin{aligned}
\max_{ij}|s_{ij}(\widehat{\sigma}_{u,ij})-\sigma_{u,ij}| &\le \max_{ij}|s_{ij}(\widehat{\sigma}_{u,ij})-\widehat{\sigma}_{u,ij}| + |\widehat{\sigma}_{u,ij}-\sigma_{u,ij}| \\
&\le \max_{ij}|s_{ij}(\widehat{\sigma}_{u,ij})-\widehat{\sigma}_{u,ij}| + O_P(w_1) \\
&= O_P(w_1).
\end{aligned}$$

where $s_{ij}(\cdot)$ is the hard thresholding at the level defined in step ii. of Method 1. Hence, $\|\widehat{\mathbf{\Sigma}}_{u,S}^{(1)} - \mathbf{\Sigma}_{u,S}\|_{\max} = O_P(w_1)$. Similarly, $\|\widehat{\mathbf{\Sigma}}_{u,S}^{(2)} - \mathbf{\Sigma}_{u,S}\|_{\max} = O_P(w_2)$. For the oracle estimator,

$$\begin{aligned}
\|\widehat{\mathbf{B}}_o\widehat{\mathbf{B}}_o' - \mathbf{B}\mathbf{B}'\|_{\max} &= \max_{ij}\left(|(\widehat{\mathbf{b}}_i^o-\mathbf{b}_i)'(\widehat{\mathbf{b}}_i-\mathbf{b}_i)| + 2|(\widehat{\mathbf{b}}_i^o-\mathbf{b}_i)'\mathbf{b}_j|\right) \\
&\le \left(\max_i\|\widehat{\mathbf{b}}_i^o-\mathbf{H}_1\mathbf{b}_i\|\right)^2 + 2\max_{ij}\|\widehat{\mathbf{b}}_i^o-\mathbf{b}_i\|\|\mathbf{b}_j\| \\
&= O_P(w_o),
\end{aligned}$$

where the last equality follows from condition (vi) and Lemma 2. Using similar arguments as in the above, $\max_{ij}|\widehat{\sigma}^o_{u,ij}-\sigma_{u,ij}| = O_P(w_0)$. Hence, $\|\widehat{\mathbf{\Sigma}}^o_{u,S}-\mathbf{\Sigma}_{u,S}\|_{\max} = O_P(w_o)$.

(3) For Method 1, let $\tilde{\mathbf{\Sigma}}_S = \mathbf{B}_S\mathbf{H}_1'\mathbf{H}_1\mathbf{B}_S' + \mathbf{\Sigma}_{u,S}$. We have

$$\|(\widehat{\mathbf{\Sigma}}^{(1)}_S)^{-1} - \mathbf{\Sigma}_S^{-1}\| \le \|(\widehat{\mathbf{\Sigma}}^{(1)}_S)^{-1} - \tilde{\mathbf{\Sigma}}_S^{-1}\| + \|\tilde{\mathbf{\Sigma}}_S^{-1} - \mathbf{\Sigma}_S^{-1}\|.$$

Since $\widehat{\mathbf{\Sigma}}^{(1)}_S = \widehat{\mathbf{B}}_1\widehat{\mathbf{B}}_1' + \widehat{\mathbf{\Sigma}}^{(1)}_{u,S}$, by Sherman-Morrison-Woodbury formula,

$$\tilde{\mathbf{\Sigma}}_S^{-1} = \mathbf{\Sigma}_{u,S}^{-1} + \mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1'\mathbf{G}^{-1}\mathbf{H}_1\mathbf{B}_S\mathbf{\Sigma}_{u,S}^{-1},$$
$$(\widehat{\mathbf{\Sigma}}^{(1)}_S)^{-1} = (\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1} + (\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1}\widehat{\mathbf{B}}_1\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1},$$

where $\mathbf{G} = \mathbf{I}_K + \mathbf{H}_1\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1'$ and $\widehat{\mathbf{G}} = \mathbf{I}_K + \widehat{\mathbf{B}}_1'(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1}\widehat{\mathbf{B}}_1$. Therefore, $\|(\widehat{\mathbf{\Sigma}}^{(1)}_S)^{-1} - \tilde{\mathbf{\Sigma}}_S^{-1}\| \le \sum_{i=1}^6 I_i$, where

$$\begin{aligned}
I_1 &= \|(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\|,\\
I_2 &= \|\{(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\}\widehat{\mathbf{B}}_1\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1}\|,\\
I_3 &= \|\{(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\}\widehat{\mathbf{B}}_1\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'\mathbf{\Sigma}_{u,S}^{-1}\|,\\
I_4 &= \|\mathbf{\Sigma}_{u,S}^{-1}(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'\mathbf{\Sigma}_{u,S}^{-1}\|,\\
I_5 &= \|\mathbf{\Sigma}_{u,S}^{-1}(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')\widehat{\mathbf{G}}^{-1}\mathbf{H}_1\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\|,\\
I_6 &= \|\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1'\{\widehat{\mathbf{G}}^{-1} - \mathbf{G}^{-1}\}\mathbf{H}_1\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\|.
\end{aligned}$$

From Lemma 3, $I_1 = O_P(m_s w_1)$. For I_2, we have

$$I_2 \le \|(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\|\|\widehat{\mathbf{B}}_1\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'\|\|(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1}\|.$$

By Lemma 3 and condition (v), $\|(\widehat{\mathbf{\Sigma}}^{(1)}_{u,S})^{-1}\| = O_P(1)$. Lemma A.6(ii) implies that $\|\widehat{\mathbf{G}}^{-1}\| = O_P(s^{-1})$. Therefore, $\|\widehat{\mathbf{B}}_1\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'\| = O_P(1)$ and $I_2 = O_P(m_s w_1)$. Similarly, $I_3 = O_P(m_s w_1)$. For I_4, condition (v) implies that $\|\mathbf{\Sigma}_{u,S}^{-1}\| = O(1)$. Next, $\|(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'\|$ is bounded by

$$\|(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'\| \le \|(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')\widehat{\mathbf{G}}^{-1}(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')'\|^{1/2}\|\widehat{\mathbf{B}}_1\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'\|^{1/2}.$$

Since $\|\widehat{\mathbf{G}}^{-1}\| = O_P(s^{-1})$ by Lemma A.6(ii) and $\|\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1'\|_F^2 = O_P(sw_1^2)$ by Lemma A.4(i), we have $\|(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')\widehat{\mathbf{G}}^{-1}(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')'\| = O_P(w_1^2)$. This together with $\|\widehat{\mathbf{B}}_1\widehat{\mathbf{G}}^{-1}\widehat{\mathbf{B}}_1'\|$ $= O_P(1)$ imply that $I_4 = O_P(w_1)$. Similarly, $I_5 = O_P(w_1)$. For I_6, we have

$$I_6 \le \|\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1'\mathbf{H}_1\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\|\|\widehat{\mathbf{G}}^{-1} - \mathbf{G}^{-1}\|.$$

Condition (ii), (v) and $\|\mathbf{H}_1\| = O_P(1)$ imply that $\|\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1'\mathbf{H}_1\mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\| = O_P(s)$. Next, we bound $\|\widehat{\mathbf{G}}^{-1} - \mathbf{G}^{-1}\|$. Note that,

$$\begin{aligned}\|\widehat{\mathbf{G}}^{-1} - \mathbf{G}^{-1}\| = \|\mathbf{G}^{-1}(\widehat{\mathbf{G}} - \mathbf{G})\widehat{\mathbf{G}}^{-1}\| &= O_P\left(s^{-2}\|\widehat{\mathbf{B}}_1'(\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)})^{-1}\widehat{\mathbf{B}}_1 - (\mathbf{B}_S\mathbf{H}_1')'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1'\|\right) \\ &= O_P\left(s^{-1}m_s w_1\right),\end{aligned}$$

because by Lemma A.6 (i) and (ii), $\|\mathbf{G}^{-1}\| = O(s^{-1})$, $\|\widehat{\mathbf{G}}^{-1}\| = O_P(s^{-1})$, and

$$\begin{aligned}&\|\widehat{\mathbf{B}}_1'(\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)})^{-1}\widehat{\mathbf{B}}_1 - (\mathbf{B}_S\mathbf{H}_1')'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1'\| \\ &\leq \|(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')'(\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)})^{-1}(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')\| + 2\|(\widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1')(\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)})^{-1}\mathbf{B}_S\mathbf{H}_1'\| \\ &\quad + \|(\mathbf{B}_S\mathbf{H}_1')'\{(\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)})^{-1} - \boldsymbol{\Sigma}_{u,S}^{-1}\}\mathbf{B}_S\mathbf{H}_1'\| \\ &= O_P\left(sw_1^2\right) + O_P(sw_1) + O_P(sm_s w_1) \\ &= O_P(sm_s w_1). \end{aligned} \tag{A.6}$$

Therefore, $I_6 = O_P(m_s w_1)$. Summing the six terms, we have $\|(\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)})^{-1} - \tilde{\boldsymbol{\Sigma}}_S^{-1}\| = O_P(m_s w_1)$. Next, we bound $\|\tilde{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\|$.

By using Sherman-Morrison-Woodbury formula again,

$$\begin{aligned}\|\tilde{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1}\| &= \left\|\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S\{[(\mathbf{H}_1'\mathbf{H}_1)^{-1} + \mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S]^{-1} - [\mathbf{I}_K + \mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S]^{-1}\}\mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\right\| \\ &= O(s)\left\|[(\mathbf{H}_1'\mathbf{H}_1)^{-1} + \mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S]^{-1} - [\mathbf{I}_K + \mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S]^{-1}\right\| \\ &= O_P\left(s^{-1}\right)\|(\mathbf{H}_1'\mathbf{H}_1)^{-1} - \mathbf{I}_K\| \\ &= o_P(m_s w_1).\end{aligned}$$

Therefore, $\|(\widehat{\boldsymbol{\Sigma}}_{u,S}^{(1)})^{-1} - \boldsymbol{\Sigma}_S^{-1}\| = O_P(m_s w_1)$. A similar result can be shown that $\|(\widehat{\boldsymbol{\Sigma}}_{u,S}^{(2)})^{-1} - \boldsymbol{\Sigma}_S^{-1}\| = O_P(m_s w_2)$.

For the oracle estimator, by Sherman-Morrison-Woodbury formula, $\|(\widehat{\boldsymbol{\Sigma}}_S^{o})^{-1} - \boldsymbol{\Sigma}_S^{-1}\| \leq \sum_{i=1}^{6} I_i$, where

$$\begin{aligned}I_1 &= \|(\widehat{\boldsymbol{\Sigma}}_{u,S}^{o})^{-1} - \boldsymbol{\Sigma}_{u,S}^{-1}\|, \\ I_2 &= \|\{(\widehat{\boldsymbol{\Sigma}}_{u,S}^{o})^{-1} - \boldsymbol{\Sigma}_{u,S}^{-1}\}\widehat{\mathbf{B}}_o\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{B}}_o'(\widehat{\boldsymbol{\Sigma}}_{u,S}^{o})^{-1}\|, \\ I_3 &= \|\{(\widehat{\boldsymbol{\Sigma}}_{u,S}^{o})^{-1} - \boldsymbol{\Sigma}_{u,S}^{-1}\}\widehat{\mathbf{B}}_o\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{B}}_o'\boldsymbol{\Sigma}_{u,S}^{-1}\|, \\ I_4 &= \|\boldsymbol{\Sigma}_{u,S}^{-1}(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{B}}_o'\boldsymbol{\Sigma}_{u,S}^{-1}\|,\end{aligned}$$

$$I_5 = \|\boldsymbol{\Sigma}_{u,S}^{-1}(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\widehat{\mathbf{J}}^{-1}\mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\|,$$
$$I_6 = \|\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S\{\widehat{\mathbf{J}}^{-1} - \mathbf{J}^{-1}\}\mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\|,$$

that $\widehat{\mathbf{J}} = \mathbf{I}_K + \widehat{\mathbf{B}}_o'(\widehat{\boldsymbol{\Sigma}}_{u,S}^o)^{-1}\widehat{\mathbf{B}}_o$ and $\mathbf{J} = \mathbf{I}_K + \mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S$.

By Lemma 3, $I_1 = O_P(m_s w_o)$. For I_2, Lemma A.6(ii) implies that $\|\widehat{\mathbf{J}}^{-1}\| = O_P(s^{-1})$. This together with condition (ii) imply that $\|\widehat{\mathbf{B}}_o\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{B}}_o'\| = O_P(1)$. Moreover, it follows from Lemma 3 and condition (v) that $\|(\widehat{\boldsymbol{\Sigma}}_{u,S}^o)^{-1}\| = O_P(1)$. Therefore,

$$I_2 \le \|(\widehat{\boldsymbol{\Sigma}}_{u,S}^o)^{-1} - \boldsymbol{\Sigma}_{u,S}^{-1}\|\|\widehat{\mathbf{B}}_o\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{B}}_o'\|\|(\widehat{\boldsymbol{\Sigma}}_{u,S}^o)^{-1}\| = O_P(m_s w_o).$$

Similarly, $I_3 = O_P(m_s w_o)$. For I_4, we have $I_4 \le \|(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\widehat{\mathbf{J}}^{-1}\mathbf{B}_S'\|\|\boldsymbol{\Sigma}_{u,S}^{-1}\|^2$. We bound $\|(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\widehat{\mathbf{J}}^{-1}\mathbf{B}_S'\|$ by

$$\|(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\widehat{\mathbf{J}}^{-1}\mathbf{B}_S'\| \le \|(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\widehat{\mathbf{J}}^{-1}(\widehat{\mathbf{B}}_o - \mathbf{B}_S)'\|^{1/2}\|\mathbf{B}_S\widehat{\mathbf{J}}^{-1}\mathbf{B}_S'\|^{1/2}.$$

Since $\|(\widehat{\mathbf{B}}_o - \mathbf{B}_S)(\widehat{\mathbf{B}}_o - \mathbf{B}_S)'\| \le \|\widehat{\mathbf{B}}_o - \mathbf{B}_S\|_F^2 \le s(\max_s\|\widehat{\mathbf{b}}_i^o - \mathbf{b}_i\|)^2 = O_P(sw_o^2)$. This together with $\|\widehat{\mathbf{J}}^{-1}\| = O_P(s^{-1})$ and $\|\widehat{\mathbf{B}}_o\widehat{\mathbf{J}}^{-1}\widehat{\mathbf{B}}_o\| = O_P(1)$ imply that $I_4 = O_P(w_o)$. Similarly, $I_5 = O_P(w_o)$. For I_6, we have $I_6 \le \|\widehat{\mathbf{J}}^{-1} - \mathbf{J}^{-1}\|\|\boldsymbol{\Sigma}_{u,S}^{-1}\|^2\|\mathbf{B}_S\mathbf{B}_S'\|$. By conditions (ii) and (iv), we have $\|\boldsymbol{\Sigma}_{u,S}^{-1}\| = O(1)$ and $\|\mathbf{B}_S\mathbf{B}_S'\| = O(s)$. As for $\|\widehat{\mathbf{J}}^{-1} - \mathbf{J}^{-1}\|$, we have

$$\|\widehat{\mathbf{J}}^{-1} - \mathbf{J}^{-1}\| = \|\widehat{\mathbf{J}}^{-1}(\widehat{\mathbf{J}} - \mathbf{J})\mathbf{J}^{-1}\| = O_P\left(s^{-2}\|\mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S - \widehat{\mathbf{B}}_o'\widehat{\boldsymbol{\Sigma}}_{u,S}^{-1}\widehat{\mathbf{B}}_o\|\right) = O_P\left(s^{-1}m_s w_o\right),$$

where the last equation follows from that

$$\begin{aligned}
\|\widehat{\mathbf{B}}_o'\widehat{\boldsymbol{\Sigma}}_{u,S}^{-1}\widehat{\mathbf{B}}_o - \mathbf{B}_S'\boldsymbol{\Sigma}_{u,S}^{-1}\mathbf{B}_S\| &\le \|(\widehat{\mathbf{B}}_o - \mathbf{B}_S)'\widehat{\boldsymbol{\Sigma}}_{u,S}^{-1}(\widehat{\mathbf{B}}_o - \mathbf{B}_S)\| + 2\|(\widehat{\mathbf{B}_o} - \mathbf{B}_S)'\widehat{\boldsymbol{\Sigma}}_{u,S}^{-1}\mathbf{B}_S\| \\
&\quad + \|\mathbf{B}_S'\{(\widehat{\boldsymbol{\Sigma}}_{u,S}^o)^{-1} - \boldsymbol{\Sigma}_{u,S}^{-1}\}\mathbf{B}_S\| \\
&= O_P\left(sw_o^2\right) + O_P(sw_o) + O_P(sm_s w_o) \\
&= O_P(sm_s w_o).
\end{aligned}$$

Therefore, $I_6 = O_P(m_s w_o)$. After summing up, $\|(\widehat{\boldsymbol{\Sigma}}_S^o)^{-1} - \boldsymbol{\Sigma}_S^{-1}\| = O_P(m_s w_o)$. □

A.3 Convergence Rates of $\bar{\boldsymbol{\Sigma}}_S$ in Section 5

Let $\bar{\mathbf{H}} = M^{-1}\sum_{m=1}^{M}\mathbf{H}_{[m]}$, where $\mathbf{H}_{[m]} = \widehat{\mathbf{V}}_m^{-1}\widehat{\mathbf{F}}_m'\mathbf{F}_m\mathbf{B}_m'\tilde{\boldsymbol{\Sigma}}_{u,m}^{-1}\mathbf{B}_m/T$, $\widehat{\mathbf{V}}_m$ is the diagonal matrix of the K largest eigenvalues of $\mathbf{Y}_m'\tilde{\boldsymbol{\Sigma}}_{u,m}^{-1}\mathbf{Y}_m/T$, $\mathbf{B}_m$ and $\mathbf{F}_m$ are the loadings and the factors in the mth group.

According to the proof of Theorem 1, the key is to show that $\max_{1\le t\le T}\|\bar{\mathbf{f}}_t - \bar{\mathbf{H}}\mathbf{f}_t\|$ has the same rate as $\max_{1\le t\le T}\|\widehat{\mathbf{f}}_t^{(2)} - \mathbf{H}_2\mathbf{f}_t\|$ and $\max_{i\le s}\|\bar{\mathbf{b}}_i - \bar{\mathbf{H}}\mathbf{b}_i\|$ has the same rate as $\max_{1\le i\le s}\|\widehat{\mathbf{b}}_i^{(2)} - \mathbf{H}_2\mathbf{b}_i\|$.

To give the rate of $\max_{1\le t\le T}\|\bar{\mathbf{f}}_t - \bar{\mathbf{H}}\mathbf{f}_t\|$, since M is fixed, p/M is in the same order as p. Then, it follows from Lemma 1 that for any $1 \le m \le M$, $\max_{1\le t\le T}\|\widehat{\mathbf{f}}_{m,t} - \mathbf{H}_{[m]}\mathbf{f}_t\| = O_P(a_{p,T})$, where $a_{p,T} = T^{-1/2} + T^{1/4}p^{-1/2}$. By definition, there exists a positive constant $C_{m,\epsilon}$ such that

$$P\left(\max_{1\le t\le T}\|\widehat{\mathbf{f}}_{m,t} - \mathbf{H}_{[m]}\mathbf{f}_t\| > C_{m,\epsilon}a_{p,T}\right) \le \epsilon/M.$$

Let $C = \max_{1\le m\le M} C_{m,\epsilon}$. We have

$$\begin{aligned}
P\left(\max_{1\le t\le T}\|\bar{\mathbf{f}}_t - \bar{\mathbf{H}}\mathbf{f}_t\| > Ca_{p,T}\right) &= P\left(\max_{1\le t\le T}\left\|\frac{1}{M}\sum_{m=1}^{M}(\widehat{\mathbf{f}}_{m,t} - \mathbf{H}_{[m]}\mathbf{f}_t)\right\| > Ca_{p,T}\right) \\
&\le \sum_{m=1}^{M} P\left(\max_{1\le t\le T}\|\widehat{\mathbf{f}}_{m,t} - \mathbf{H}_{[m]}\mathbf{f}_t\| > Ca_{p,T}\right) \\
&\le \epsilon.
\end{aligned}$$

By definition, $\max_{1\le t\le T}\|\bar{\mathbf{f}}_t - \bar{\mathbf{H}}\mathbf{f}_t\| = O_P\left(a_{p,T}\right)$, which is the same as $\max_{1\le t\le T}\|\widehat{\mathbf{f}}_t^{(2)} - \mathbf{H}_2\mathbf{f}_t\|$ shown in Lemma 1.

Next, we show that $\max_{i\le s}\|\bar{\mathbf{b}}_i - \bar{\mathbf{H}}\mathbf{b}_i\| = O_P(w_2)$. For any $1 \le m \le M$, similarly as in Lemma A.2, we have $\|\mathbf{H}_{[m]}\| = O_P(1)$. By the same union bound argument, we have $\|\bar{\mathbf{H}}\| = O_P(1)$. Then, it follows from the same proof of Lemma 2 that $\max_{i\le s}\|\bar{\mathbf{b}}_i - \bar{\mathbf{H}}\mathbf{b}_i\| = O_P(w_2)$.

As M is fixed, the results in Lemma 3 and Theorem 1 for each individual group hold. Repeatedly using the above union bound argument, $\bar{\boldsymbol{\Sigma}}_S$ is shown to have the same convergence rate as $\widehat{\boldsymbol{\Sigma}}_S^{(2)}$.

A.4 Additional Lemmas

Lemma A.1. *Under conditions of Lemma 1, it holds that*

$$\max_{i\le s,t\le T}\|(\widehat{\mathbf{b}}_i^{(1)})'\widehat{\mathbf{f}}_t^{(1)} - \mathbf{b}_i'\mathbf{f}_t\| = O_P\left((\log T)^{1/2}\sqrt{\frac{\log s}{T}} + \frac{T^{1/4}}{\sqrt{s}}\right)$$

$$\max_{i\le s,t\le T}\|(\widehat{\mathbf{b}}_i^{(2)})'\widehat{\mathbf{f}}_t^{(2)} - \mathbf{b}_i'\mathbf{f}_t\| = O_P\left((\log T)^{1/2}\sqrt{\frac{\log s}{T}} + \frac{T^{1/4}}{\sqrt{p}}\right)$$
$$\max_{i\le s,t\le T}\|(\widehat{\mathbf{b}}_i^{o})'\mathbf{f}_t - \mathbf{b}_i'\mathbf{f}_t\| = O_P\left((\log T)^{1/2}\sqrt{\frac{\log s}{T}}\right).$$

Proof of Lemma A.1. Under condition (i), it follows from the union bound argument that

$$\max_{t\le T}\|\mathbf{f}_t\| = O_P\left(\sqrt{\log T}\right).$$

Then, for Method 1, it follows from Lemmas 1, 2, A.2, and condition (vi) that, uniformly in i and t,

$$\begin{aligned}\|(\widehat{\mathbf{b}}_i^{(1)})'\widehat{\mathbf{f}}_t^{(1)} - \mathbf{b}_i'\mathbf{f}_t\| &\le \|\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}_1\mathbf{b}_i\|\|\widehat{\mathbf{f}}_t^{(1)} - \mathbf{H}_1\mathbf{f}_t\| + \|\mathbf{H}_1\mathbf{b}_i\|\|\widehat{\mathbf{f}}_t^{(1)} - \mathbf{H}_1\mathbf{f}_t\| \\ &\quad + \|\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}_1\mathbf{b}_i\|\|\mathbf{H}_1\mathbf{f}_t\| + \|\mathbf{b}_i\|\|\mathbf{f}_t\|\|\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K\|_F \\ &= O_P\left((\log T)^{1/2}\sqrt{\frac{\log s}{T}} + \frac{T^{1/4}}{\sqrt{s}}\right).\end{aligned}$$

For Method 2, similar arguments give

$$\max_{i\le s,t\le T}\|(\widehat{\mathbf{b}}_i^{(2)})'\widehat{\mathbf{f}}_t^{(2)} - \mathbf{b}_i'\mathbf{f}_t\| = O_P\left((\log T)^{1/2}\sqrt{\frac{\log s}{T}} + \frac{T^{1/4}}{\sqrt{p}}\right).$$

In the oracle setting, where the factors are known, we have

$$\begin{aligned}\max_{i\le s,t\le T}\|(\widehat{\mathbf{b}}_i^{o})'\mathbf{f}_t - \mathbf{b}_i'\mathbf{f}_t\| = \max_{i\le s,t\le T}\|\widehat{\mathbf{b}}_i^{o} - \mathbf{b}_i\|\|\mathbf{f}_t\| &= O_P\left(\sqrt{\log T}\max_{i\le s}\|\widehat{\mathbf{b}}_i^{o} - \mathbf{b}_i\|\right) \\ &= O_P\left((\log T)^{1/2}\sqrt{\frac{\log s}{T}}\right).\end{aligned}$$

□

Lemma A.2. *Let* $\mathbf{H}_1 = \widehat{\mathbf{V}}_1^{-1}\widehat{\mathbf{F}}^{(1)\prime}\mathbf{F}\mathbf{B}_S'\tilde{\mathbf{\Sigma}}_{u,S}^{-1}\mathbf{B}_S/T$ *and* $\mathbf{H}_2 = \widehat{\mathbf{V}}_2^{-1}\widehat{\mathbf{F}}^{(2)\prime}\mathbf{F}\mathbf{B}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{B}/T$*, where* $\widehat{\mathbf{V}}_1$ *is the diagonal matrix of the largest* K *eigenvalues of* $\mathbf{Y}_S'\tilde{\mathbf{\Sigma}}_{u,S}^{-1}\mathbf{Y}_S/T$ *and* $\widehat{\mathbf{V}}_2$ *is the diagonal matrix of the largest* K *eigenvalues of* $\mathbf{Y}'\tilde{\mathbf{\Sigma}}_u^{-1}\mathbf{Y}/T$*. Under conditions of Lemma 1,* $\|\mathbf{H}_1\| = O_P(1)$ *and* $\|\mathbf{H}_2\| = O_P(1)$*.*

Proof of Lemma A.2. Since $\mathbf{\Sigma}_{u,S}$ is a submatrix of $\mathbf{\Sigma}_u$, it follows from condition (v) that $\lambda_{\min}(\mathbf{\Sigma}_{u,S}^{-1}) \ge c_2^{-1}$. By Proposition 4.1 of Bai and Liao (2013), $\|\tilde{\mathbf{\Sigma}}_{u,S}^{-1} - \mathbf{\Sigma}_{u,S}^{-1}\| = o_P(1)$. Therefore, with probability tending to 1, $\|\tilde{\mathbf{\Sigma}}_{u,S}^{-1}\| \ge 1/(2c_2)$. Then,

$$T^{-1}\mathbf{Y}_S'\tilde{\mathbf{\Sigma}}_{u,S}^{-1}\mathbf{Y}_S = T^{-1}\mathbf{Y}_S'(\tilde{\mathbf{\Sigma}}_{u,S}^{-1} - (1/2c_2)\mathbf{I})\mathbf{Y}_S + 1/(2c_2T)\mathbf{Y}_S'\mathbf{Y}_S.$$

Under the pervasive condition (i), it follows from Lemma C.4 of Fan et al. (2013) that the Kth largest eigenvalue of $T^{-1}\mathbf{Y}_S'\mathbf{Y}_S$ is larger than Ms. Since $T^{-1}\mathbf{Y}_S'(\tilde{\boldsymbol{\Sigma}}_{u,S}^{-1} - (1/2c_2)\mathbf{I})\mathbf{Y}_S$ is semi-positive definite, it follows from Weyl's inequality that

$$\lambda_K(T^{-1}\mathbf{Y}_S'\tilde{\boldsymbol{\Sigma}}_{u,S}^{-1}\mathbf{Y}_S) \geq \lambda_K(1/(2c_2T)\mathbf{Y}_S'\mathbf{Y}_S) \geq Ms/(2c_2).$$

Hence $||\widehat{\mathbf{V}}_1^{-1}|| = O_P(s^{-1})$. Also, $\lambda_{\max}(\|\mathbf{F}'\mathbf{F}\|) = \lambda_{\max}(\|\sum_{t=1}^T \mathbf{f}_t\mathbf{f}_t'\|) = O_P(T)$. In addition, $\lambda_{\max}(\|\sum_{t=1}^T \widehat{\mathbf{f}}_t^{(1)}(\widehat{\mathbf{f}}_t^{(1)})'\|) = O_P(T)$, where the last equation follows from the constraint in (6). Then, $\|(\widehat{\mathbf{F}}^{(1)})'\mathbf{F}\| \leq \|(\widehat{\mathbf{F}}^{(1)})'\widehat{\mathbf{F}}^{(1)}\|^{1/2}\|\mathbf{F}'\mathbf{F}\|^{1/2} = O_P(T)$. These results together with $\|\mathbf{B}_S'\tilde{\boldsymbol{\Sigma}}_{u,S}^{-1}\mathbf{B}_S\| = O(s)$ imply that $\|\mathbf{H}_1\| = O_P(1)$. Similarly, $\|\mathbf{H}_2\| = O_P(1)$. □

Lemma A.3. *(i)* $\|\mathbf{H}_1\mathbf{H}_1' - \mathbf{I}_K\|_F = O_P\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{s}}\right)$*; (ii)* $\|\mathbf{H}_2\mathbf{H}_2' - \mathbf{I}_K\|_F = O_P\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}}\right)$. *(iii)* $\|\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K\|_F = O_P\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{s}}\right)$*; (iv)* $\|\mathbf{H}_2'\mathbf{H}_2 - \mathbf{I}_K\|_F = O_P\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}}\right)$.

Proof of Lemma A.3. Let $\widehat{\mathrm{cov}}(\mathbf{H}_1\mathbf{f}_t) = \frac{1}{T}\sum_{t=1}^T(\mathbf{H}_1\mathbf{f}_t)(\mathbf{H}_1\mathbf{f}_t)'$. Then,

$$\|\mathbf{H}_1\mathbf{H}_1' - \mathbf{I}_K\|_F \leq \underbrace{\|\mathbf{H}_1\mathbf{H}_1' - \widehat{\mathrm{cov}}(\mathbf{H}_1\mathbf{f}_t)\|_F}_{I_1} + \underbrace{\|\widehat{\mathrm{cov}}(\mathbf{H}_1\mathbf{f}_t) - \mathbf{I}_K\|_F}_{I_2}.$$

For I_1, we have $I_1 \leq \|\mathbf{H}_1\|^2\|\mathbf{I}_K - \widehat{\mathrm{cov}}(\mathbf{f}_t)\|_F$, where $\widehat{\mathrm{cov}}(\mathbf{f}_t) = \frac{1}{T}\sum_{t=1}^T \mathbf{f}_t\mathbf{f}_t'$. It follows from Lemma C.3(i) of Fan et al. (2013) that $\|\mathbf{I}_K - \widehat{\mathrm{cov}}(\mathbf{f}_t)\|_F = O_P\left(1/\sqrt{T}\right)$. Then, $I_1 = O_P\left(1/\sqrt{T}\right)$, since $\|\mathbf{H}_1\| = O_P(1)$. For I_2, by the identifiability constraint in (6), $\frac{1}{T}\sum_{t=1}^T \widehat{\mathbf{f}}_t^{(1)}\widehat{\mathbf{f}}_t^{(1)\prime} = \mathbf{I}_K$. Therefore,

$$\begin{aligned}
I_2 &= \left\|\frac{1}{T}\sum_{t=1}^T \mathbf{H}_1\mathbf{f}_t(\mathbf{H}_1\mathbf{f}_t)' - \frac{1}{T}\sum_{t=1}^T \widehat{\mathbf{f}}_t^{(1)}\widehat{\mathbf{f}}_t^{(1)\prime}\right\|_F \\
&\leq \left\|\frac{1}{T}\sum_{t=1}^T(\mathbf{H}_1\mathbf{f}_t - \widehat{\mathbf{f}}_t^{(1)})(\mathbf{H}_1\mathbf{f}_t)'\right\|_F + \left\|\frac{1}{T}\sum_{t=1}^T \widehat{\mathbf{f}}_t^{(1)}(\widehat{\mathbf{f}}_t^{(1)} - \mathbf{H}_1\mathbf{f}_t)'\right\|_F \\
&\leq \left(\frac{1}{T}\sum_{t=1}^T\|\mathbf{H}_1\mathbf{f}_t - \widehat{\mathbf{f}}_t^{(1)}\|^2 \cdot \frac{1}{T}\sum_{t=1}^T\|\mathbf{H}_1\mathbf{f}_t\|^2\right)^{1/2} + \left(\frac{1}{T}\sum_{t=1}^T\|\mathbf{H}_1\mathbf{f}_t - \widehat{\mathbf{f}}_t^{(1)}\|^2 \cdot \frac{1}{T}\sum_{t=1}^T\|\widehat{\mathbf{f}}_t^{(1)}\|^2\right)^{1/2} \\
&= O_P\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{s}}\right),
\end{aligned}$$

where the last equality follows from Lemma A.5 and that $\|\mathbf{H}_1\mathbf{f}_t\| \leq \|\mathbf{H}_1\|\|\mathbf{f}_t\| = O_P(1)$ and $\|\widehat{\mathbf{f}}_t^{(1)}\| = O_P(1)$. Similarly, $\|\mathbf{H}_2\mathbf{H}_2' - \mathbf{I}_K\|_F = O_P\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{p}}\right)$.

(iii) Since $\|\mathbf{H}_1\mathbf{H}_1' - \mathbf{I}_K\|_F = O_P\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{s}}\right)$ and $\|\mathbf{H}_1\| = O_P(1)$, we have $\|\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_1 - \mathbf{H}_1\|_F = O_P\left(\frac{1}{\sqrt{T}} + \frac{1}{\sqrt{s}}\right)$. Since $\mathbf{H}_1^{-1} = \mathbf{H}_1^{-1}(\mathbf{I}_K - \mathbf{H}_1\mathbf{H}_1' + \mathbf{H}_1\mathbf{H}_1')$, it follows Lemma A.3(i) that $\|\mathbf{H}_1^{-1}\| \leq \|\mathbf{H}_1^{-1}\| O_P\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{s}}\right) + \|\mathbf{H}_1'\|$. Hence, $\|\mathbf{H}_1^{-1}\| = O_P(1)$. Left multiplying $\mathbf{H}_1\mathbf{H}_1'\mathbf{H}_1 - \mathbf{H}_1$ by $\mathbf{H}_1^{-1}$ gives $\|\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K\|_F = O_P\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{s}}\right)$. Similarly, $\|\mathbf{H}_2'\mathbf{H}_2 - \mathbf{I}_K\|_F = O_P\left(\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{p}}\right)$. □

Lemma A.4. *Let $\mathbf{C}_1 = \widehat{\mathbf{B}}_1 - \mathbf{B}_S\mathbf{H}_1'$ and $\mathbf{C}_2 = \widehat{\mathbf{B}}_2 - \mathbf{B}_S\mathbf{H}_2'$, where $\widehat{\mathbf{B}}_1$, $\widehat{\mathbf{B}}_2$, and $\mathbf{B}_S$ are defined in Section 4.*

(i) $\|\mathbf{C}_1\|_F^2 = O_P(sw_1^2)$, $\|\mathbf{C}_2\|_F^2 = O_P(sw_2^2)$; $\|\mathbf{C}_1\mathbf{C}_1'\|_{\mathbf{\Sigma}_S}^2 = O_P(sw_1^4)$, $\|\mathbf{C}_2\mathbf{C}_2'\|_{\mathbf{\Sigma}_S}^2 = O_P(sw_2^4)$.

(ii) $\|\widehat{\mathbf{\Sigma}}_{u,S}^{(1)} - \mathbf{\Sigma}_{u,S}\|_{\mathbf{\Sigma}_S}^2 = O_P(m_s^2 w_1^2)$; $\|\widehat{\mathbf{\Sigma}}_{u,S}^{(2)} - \mathbf{\Sigma}_{u,S}\|_{\mathbf{\Sigma}_S}^2 = O_P(m_s^2 w_2^2)$.

(iii) $\|\mathbf{B}_S\mathbf{H}_1'\mathbf{C}_1'\|_{\mathbf{\Sigma}_S}^2 = O_P(w_1^2)$; $\|\mathbf{B}_S\mathbf{H}_2'\mathbf{C}_2'\|_{\mathbf{\Sigma}_S}^2 = O_P(w_2^2)$.

(iv) $\|\mathbf{B}_S(\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K)\mathbf{B}_S'\|_{\mathbf{\Sigma}_S}^2 = O_P\left(\frac{1}{sT}+\frac{1}{s^2}\right)$; $\|\mathbf{B}_S(\mathbf{H}_2'\mathbf{H}_2 - \mathbf{I}_K)\mathbf{B}_S'\|_{\mathbf{\Sigma}_S}^2 = O_P\left(\frac{1}{sT}+\frac{1}{sp}\right)$.

Proof of Lemma A.4. (i) We have $\|\mathbf{C}_1\|_F^2 \leq s(\max_{i\leq s}\|\widehat{\mathbf{b}}_i^{(1)} - \mathbf{H}\mathbf{b}_i\|)^2 = O_P(sw_1^2)$. By the general result that for any matrix $\mathbf{A}$, $\|\mathbf{A}\|_{\mathbf{\Sigma}_S}^2 = s^{-1}\|\mathbf{\Sigma}_S^{-1/2}\mathbf{A}\mathbf{\Sigma}_S^{-1/2}\|_F^2 = O_P(s^{-1}\|\mathbf{A}\|_F^2)$, we have $\|\mathbf{C}_1'\mathbf{C}_1\|_{\mathbf{\Sigma}_S}^2 = O_P(s^{-1}\|\mathbf{C}_1\|_F^4) = O_P(sw_1^4)$. Similarly, $\|\mathbf{C}_2\|_F^2 = O_P(sw_2^2)$ and $\|\mathbf{C}_2\mathbf{C}_2'\|_{\mathbf{\Sigma}_S}^2 = O_P(sw_2^4)$.

(ii) By Lemma 3,

$$\|\widehat{\mathbf{\Sigma}}_{u,S}^{(1)} - \mathbf{\Sigma}_{u,S}\|_{\mathbf{\Sigma}_S}^2 = O_P\left(s^{-1}\|\widehat{\mathbf{\Sigma}}_{u,S}^{(1)} - \mathbf{\Sigma}_{u,S}\|_F^2\right) = O_P\left(\|\widehat{\mathbf{\Sigma}}_{u,S}^{(1)} - \mathbf{\Sigma}_{u,S}\|^2\right) = O_P\left(m_s^2 w_1^2\right).$$

Similar results can be shown for $\|\widehat{\mathbf{\Sigma}}_{u,S}^{(2)} - \mathbf{\Sigma}_{u,S}\|_{\mathbf{\Sigma}_S}$.

(iii) By adapt the proof of Theorem 2 in Fan et al. (2008), we have that $\|\mathbf{B}_S'\mathbf{\Sigma}_S^{-1}\mathbf{B}_S\| = O(1)$. Hence,

$$\begin{aligned}\|\mathbf{B}_S\mathbf{H}_1'\mathbf{C}_1'\|_{\mathbf{\Sigma}_S}^2 &= s^{-1}\mathrm{tr}(\mathbf{H}_1'\mathbf{C}_1'\mathbf{\Sigma}_S^{-1}\mathbf{C}_1\mathbf{H}_1\mathbf{B}_S'\mathbf{\Sigma}_S^{-1}\mathbf{B}_S)\\ &\leq s^{-1}\|\mathbf{H}_1\|^2\|\mathbf{B}_S'\mathbf{\Sigma}_S^{-1}\mathbf{B}_S\|\|\mathbf{\Sigma}_S^{-1}\|\|\mathbf{C}_1\|_F^2\\ &= O_P\left(s^{-1}\|\mathbf{C}_1\|_F^2\right) = O_P\left(w_1^2\right).\end{aligned}$$

Similarly, $\|\mathbf{B}_S\mathbf{H}_2'\mathbf{C}_2'\|_{\mathbf{\Sigma}_S} = O_P(w_2^2)$.

(iv) We have

$$\begin{aligned}\|\mathbf{B}_S(\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K)\mathbf{B}_S'\|_{\mathbf{\Sigma}_S}^2 &= s^{-1}\mathrm{tr}((\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K)\mathbf{B}_S'\mathbf{\Sigma}_S^{-1}\mathbf{B}_S(\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K)\mathbf{B}_S'\mathbf{\Sigma}_S^{-1}\mathbf{B}_S)\\ &\leq s^{-1}\|\mathbf{H}_1'\mathbf{H}_1 - \mathbf{I}_K\|_F^2\|\mathbf{B}_S'\mathbf{\Sigma}_S^{-1}\mathbf{B}_S\|^2 = O_P\left(\frac{1}{sT}+\frac{1}{s^2}\right).\end{aligned}$$

Similarly, $\|\mathbf{B}_S(\mathbf{H}_2'\mathbf{H}_2 - \mathbf{I}_K)\mathbf{B}_S'\|_{\mathbf{\Sigma}_S}^2 = O_P\left(\frac{1}{sT}+\frac{1}{sp}\right)$. □

Lemma A.5. *Under conditions of Lemma 1,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t^{(1)}-\mathbf{H}_1\mathbf{f}_t\|^2=O_P\left(1/s+1/T\right)$$
$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t^{(2)}-\mathbf{H}_2\mathbf{f}_t\|^2=O_P\left(1/p+1/T\right)$$

Proof of Lemma A.5. Without loss of generality, we only prove the result for general p. Again, we write $\widehat{\mathbf{f}}_t^{(2)}$ as $\widehat{\mathbf{f}}_t$, $\mathbf{H}_2$ as $\mathbf{H}$ and $\widehat{\mathbf{V}}_2$ as $\widehat{\mathbf{V}}$ for notational simplicity. By (A.2),

$$\frac{1}{T}\sum_{t=1}^{T}\|\widehat{\mathbf{f}}_t-\mathbf{H}\mathbf{f}_t\|^2\le c\|\widehat{\mathbf{V}}^{-1}\|^2\sum_{j=1}^{7}\frac{1}{T}\sum_{t=1}^{T}G_{jt}^2,$$

where c is a positive constant and G_{jt} is the jth summand on the right hand side of (A.2). By Lemma A.6 (iv) of Bai and Liao (2013), $\frac{1}{T}\sum_{i=1}^{T}G_{1t}^2=o_P\left(1/p+1/T\right)$. By Lemma A.10 (i) and (iii) of Bai and Liao (2013), $\frac{1}{T}\sum_{t=1}^{T}G_{2t}^2=O_P\left(1/T\right)$ and $\frac{1}{T}\sum_{t=1}^{T}G_{3t}^2=O_P\left(1/T\right)$. By Lemma A.6 (iii), (v) and (vi) of Bai and Liao (2013), $\frac{1}{T}\sum_{t=1}^{T}G_{4t}^2=o_P(1/p)$, $\frac{1}{T}\sum_{t=1}^{T}G_{6t}^2=o_P(1/p)$ and $\frac{1}{T}\sum_{t=1}^{T}G_{7t}^2=o_P(1/p)$. Finally, by Lemma A.11 (ii) of Bai and Liao (2013), $\frac{1}{T}\sum_{t=1}^{T}G_{5t}^2=O_P(1/p)$. Therefore, the dominating terms are G_{2t}, G_{3t} and G_{5t}, which together give the rate of $O_P\left(1/p+1/T\right)$. □

Lemma A.6. *With probability tending to 1,*
(i) $\lambda_{\min}(\mathbf{I}_K+(\mathbf{B}_S\mathbf{H}_1')'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1')\ge cs$, $\lambda_{\min}(\mathbf{I}_K+(\mathbf{B}_S\mathbf{H}_2')'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_2')\ge cs$, $\lambda_{\min}(\mathbf{I}_K+\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S)\ge cs$;
(ii) $\lambda_{\min}(\mathbf{I}_K+\widehat{\mathbf{B}}_1'(\widehat{\mathbf{\Sigma}}_{u,S}^{(1)})^{-1}\widehat{\mathbf{B}}_1)\ge cs$, $\lambda_{\min}(\mathbf{I}_K+\widehat{\mathbf{B}}_2'(\widehat{\mathbf{\Sigma}}_{u,S}^{(2)})^{-1}\widehat{\mathbf{B}}_2)\ge cs$, $\lambda_{\min}(\mathbf{I}_K+\widehat{\mathbf{B}}_o'(\widehat{\mathbf{\Sigma}}_{u,S}^{o})^{-1}\widehat{\mathbf{B}}_o)$ $\ge cs$;
(iii) $\lambda_{\min}((\mathbf{H}_1'\mathbf{H}_1)^{-1}+\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S)\ge cs$, $\lambda_{\min}((\mathbf{H}_2'\mathbf{H}_2)^{-1}+\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S)\ge cs$.

Proof of Lemma A.6. By Lemma A.3, with probability tending to one, $\lambda_{\min}(\mathbf{H}_1\mathbf{H}_1')$ is bounded away from 0. Therefore,

$$\lambda_{\min}(\mathbf{I}_K+(\mathbf{B}_S\mathbf{H}_1')'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1')\ge\lambda_{\min}(\mathbf{H}_1\mathbf{B}_S'\mathbf{\Sigma}_{u,S}^{-1}\mathbf{B}_S\mathbf{H}_1')$$
$$\ge\lambda_{\min}(\mathbf{\Sigma}_{u,S}^{-1})\lambda_{\min}(\mathbf{B}_S'\mathbf{B}_S)\lambda_{\min}(\mathbf{H}_1\mathbf{H}_1')\ge cs.$$

Similar results hold for the other two statements. The results in (ii) follow from (i) and (A.6). The statement (iii) follows from a similar argument as $\mathbf{H}_1\mathbf{H}_1'$ and $\mathbf{H}_2\mathbf{H}_2'$ are positive semi-definite. □

Lemma A.7. *[Theorem A.1 of Fan et al. (2013)] Let* $\widehat{u}_{it}$ *be defined as in step ii. of Method 1 in Section 4. Under conditions (iv), (v), if there is a sequence* $a_T = o(1)$ *so that* $\max_{i\le p} \frac{1}{T}\sum_{t=1}^{T} |u_{it} - \widehat{u}_{it}|^2 = O_P(a_T^2)$ *and* $\max_{i\le p, t\le T} |u_{it} - \widehat{u}_{it}| = o_P(1)$*, then the adaptive thresholding estimator* $\widehat{\mathbf{\Sigma}}_u$ *with* $\omega(p) = \sqrt{(\log p)/T} + a_T$ *satisfies that* $\|\widehat{\mathbf{\Sigma}}_u - \mathbf{\Sigma}_u\| = O_P(m_p[\omega(p)]^{1-q})$*. If further* $m_p[\omega(p)]^{1-q} = o(1)$*, then* $\widehat{\mathbf{\Sigma}}_u$ *is invertible with probability approaching one, and* $\|\widehat{\mathbf{\Sigma}}_u^{-1} - \mathbf{\Sigma}_u^{-1}\| = O_P(m_p[\omega(p)]^{1-q})$.

References

Ahn, S. C., and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203-1227.

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *Journal of the American Statistical Association* **96**, 939-967.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71**, 135-171.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics* **40**, 436-465.

Bai, J. and Liao, Y. (2013). Statistical inferences using large estimated covariances for panel data and factor models. *arXiv:1307.2662*.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191-221.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577-2604.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672-684.

Cai, T., Zhang, C.-H., and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* **38**, 2118-2144.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2012). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438-1456.

Choi, I. (2012). Efficient estimation of factor models. *Econometric Theory* **28**, 274-308.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**, 3-56.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147**, 186-197.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics* **39**, 3320-3356.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society B* **75**, 603-680.

Johnson, R. A. and Wichern, D. W. (1992). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ. Prentice hall.

Kapetanios, G. (2010). A testing procedure for determining the number of factors in approximate factor models with large datasets. *Journal of Business and Economic Statistics* **28**, 397-409.

Krug, T., Gabriel, J. P., Taipa, R., Fonseca, B. V., Domingues-Montanari, S., Fernandez-Cadenas, I., et al. (2012). TTC7B emerges as a novel risk factor for ischemic stroke through the convergence of several genome-wide approaches. *Journal of Cerebral Blood Flow & Metabolism* **32**, 1061-1072.

Kustra, R., Shioda, R., and Zhu, M. (2006). A factor analysis model for functional genomics. *BMC Bioinformatics* **7**, 216.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics* **37**, 4254-4278.

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

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30.

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

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* **104**, 177-186.

Shao, J. (2003). *Mathematical Statistics*. Springer-Verlag.

Stock, J. and Watson, M. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association* **97**, 1167-1179.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* **64**, 479-498.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894-942.