DOUBLE CROSS VALIDATION FOR THE NUMBER OF FACTORS IN APPROXIMATE FACTOR MODELS

XIANLI ZENG[a], YINGCUN XIA[b] AND LINJUN ZHANG[c]

Determining the number of factors is essential to factor analysis. In this paper, we propose an efficient cross validation (CV) method to determine the number of factors in approximate factor models. The method applies CV twice, first along the directions of observations and then variables, and hence is referred to hereafter as double cross-validation (DCV). Unlike most CV methods, which are prone to overfitting, the DCV is statistically consistent in determining the number of factors when both dimension of variables and sample size are sufficiently large. Simulation studies show that DCV has outstanding performance in comparison to existing methods in selecting the number of factors, especially when the idiosyncratic error has heteroscedasticity, or heavy tail, or relatively large variance.

KEYWORDS: cross validation, eigen-decomposition, factor model, high dimensional data, model selection.

1. INTRODUCTION

Factor models are a special kind of latent-variable models that are widely used in economics and other disciplines of research. Due to the ubiquitous dependence across high-dimensional economic variables, it is appealing for explaining the variation of the high-dimensional economic measurements from only a small number of latent common factors. Well-known examples include the arbitrage pricing theory (Ross, 1976), rank of demand systems (Lewbel, 1991), multiple-factor models (Fama and French, 1993), components analysis in economic activities (Gregory and Head, 1999), and diffusion index (Stock and Watson, 2002). Successful applications rely heavily on the ability to correctly specify the number of factors, which is usually unknown in practice. Therefore, determining the number

[a]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546; stazeng@nus.edu.sg

[b]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546; staxyc@nus.edu.sg

[c]Department of Statistics, the Wharton School, University of Pennsylvania, Philadelphia, PA 19104. linjunz@wharton.upenn.edu

1

of factors is an essential step in applying factor models.

One approach to this end is to utilize the eigen-structure of data matrix; see for example Fujikoshi (1977), Schott (1994), Ye and Weiss (2003), Onatski (2010) and Luo and Li (2016), amongst others. However, these methods require strong assumptions on the separation of eigenvalues, and thus most of them are only applicable to fixed dimensional data. More popular approaches are cross-validation (CV) and information criteria (IC). For factor model, several cross-validation methods have been proposed; see, for example, Wold (1978) and Eastment and Krzanowski (1982), amongst others. Bro et al. (2008) comprehensively reviewed these CV-based approaches, and concluded that most were not statistically consistent. In addition, they recommended an alternative method based on the expectation maximization (EM) algorithm, which turned out to be extremely computationally expensive. As an alternative, the ICs are more computationally effective and consistency can be guaranteed. As far as we know, Cragg and Donald (1997) was the first paper that used ICs to determine the number of factors. Subsequently, other methods based on the Akaike information criterion (AIC) and Bayesian information criterion (BIC) (Stock and Watson, 1998; Forni et al., 2000; Bai and Ng, 2002; Li et al., 2017) were proposed for different settings of sample size n and dimension of variable, p. When the variance of the idiosyncratic component is relatively small, these methods are able to provide very efficient estimates of the number of factors (Bai and Ng, 2002). However, the IC-based approaches are usually not fully data-driven since their penalties depend on predetermined tuning parameters. Even though several guides are proposed to mitigate the effect of tuning parameters, the performances of these methods are not stable when the signal-to-noise ratio is relatively small (Onatski, 2010), i.e., the variation of the idiosyncratic component is relatively large, or when the idiosyncratic component has heavy tails in distribution. Unfortunately, both of which are common in financial data.

In this paper, we propose to estimate the number of factors by a computationally-efficient CV method. Note that the conventional CV methods, e.g., those used in the linear regression, only leave one or several *observations* out. In contrast, factor models involve a matrix, where both observations (or rows) and variables (or columns) play similar roles mathematically in the analysis. Thus, we propose a double cross-validation (DCV) method that applies CV first to the *observations* and then to the *variables* in the observations. Theoretically, we show that the method is consistent for high-dimensional data and allows dependency between the idiosyncratic errors. The method is thus applicable to approximate factor models. In addition, computationally, the second CV can be

easily calculated by the linear regression error using the full data (Shao, 1993), while the first CV can be done in a K-fold manner. Simulation studies show that the proposed approach performs satisfactorily even when the idiosyncratic error has relatively big variance or homoscedasticity, or heavy tail, which are especially relevant for economic data.

The rest of this paper is organized as follows. Section 2 reviews the basic notation of factor analysis. Section 3 presents the DCV approach. Section 4 establishes statistical consistency of the approach, while the proofs of those properties are given in the Appendix. Sections 5–6 present a set of simulation studies and an empirical application to assess the finite sample performance of DCV. Concluding remarks are given in Section 7.

2. FACTOR MODEL AND ITS FACTORS AND LOADINGS

For a p-dimensional random vector $\boldsymbol{x} = (X_1, X_2, ..., X_p)^\top$, the factor model assumes the following *bilinear* structure

$$\boldsymbol{x} = \boldsymbol{L}^0 \boldsymbol{f}^0 + \boldsymbol{e},$$

where $\boldsymbol{L}^0 \in \mathbb{R}^{p \times d_0}$ is the loading matrix, $\boldsymbol{f}^0 = (f_1, f_2, ..., f_{d_0})^\top$ is the vector of d_0 common factors. Together $\boldsymbol{L}^0 \boldsymbol{f}^0$ is called the common component of the data, and $\boldsymbol{e} = (e_1, e_2, ..., e_p)^\top$ the idiosyncratic component. Let $\boldsymbol{\Sigma}_{\boldsymbol{e}} = \mathbb{E}[\boldsymbol{e}\boldsymbol{e}^\top]$, $\boldsymbol{\Sigma}_{\boldsymbol{f}_0} = \mathbb{E}[\boldsymbol{f}_0\boldsymbol{f}_0^\top]$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$. Under this model, the variation of the p-dimensional variable is mainly generated by a small number of common factors. In the statistical literature, the idiosyncratic component are usually assumed to be independent of the common component, and have diagonal covariance matrix, in which case $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{L}^0 \boldsymbol{\Sigma}_{\boldsymbol{f}_0} \boldsymbol{L}^{0\top} + \boldsymbol{\Sigma}_{\boldsymbol{e}}$. However, such independence assumption and diagonal structure are usually not realistic in financial data or macroeconomic data where factor analysis are often applied. In this paper, we consider the approximate factor model where the idiosyncratic components can be weakly dependent.

Let $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, ..., \boldsymbol{x}_n^\top)^\top$ be random samples of n observations on p variables. We have the following matrix form for the factor model:

$$(1) \qquad \boldsymbol{X} = \boldsymbol{F}^0 \boldsymbol{L}^{0\top} + \boldsymbol{E},$$

with $\boldsymbol{F}^0 = (\boldsymbol{f}_1^0, \boldsymbol{f}_2^0, ..., \boldsymbol{f}_n^0)^\top$ and $\boldsymbol{E} = (\boldsymbol{e}_1^\top, \boldsymbol{e}_2^\top, ..., \boldsymbol{e}_n^\top)^\top$. According to the eigen-decomposition, $\boldsymbol{X}^\top \boldsymbol{X}$ has the expression

$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{\Phi} \boldsymbol{\Lambda} \boldsymbol{\Phi}^\top,$$

where $\mathbf{\Lambda}$ is a $p \times p$ diagonal matrix with diagonal entries being the eigenvalues of $\mathbf{X}^\top \mathbf{X}$, $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_n) \in \mathbb{R}^{n \times n}$ is the eigenvector matrix of $\mathbf{X}^\top \mathbf{X}$ such that $\mathbf{\Phi}^\top \mathbf{\Phi} = \mathbf{I}_p$. Then, with working number of factors d, the estimators of factor loadings and factors are respectively

$$(2) \qquad \widetilde{\mathbf{L}}^d = \sqrt{p}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_d) \ \text{ and } \ \widetilde{\mathbf{F}}^d = \frac{1}{p}\mathbf{X}\widetilde{\mathbf{L}}^d.$$

Denote by $\mathrm{Tr}(\mathbf{A})$ the trace of matrix $\mathbf{A}$. It is easy to see that

$$(\widetilde{\mathbf{F}}^d, \widetilde{\mathbf{L}}^d) = \operatorname*{argmin}_{\substack{\mathbf{F}^d \in \mathbb{R}^{n \times d}, \mathbf{L}^d \in \mathbb{R}^{p \times d} \\ \mathbf{L}^{d\top}\mathbf{L}^d/p = \mathbf{I}_d}} \{\mathrm{Tr}[(\mathbf{X} - \mathbf{F}^d\mathbf{L}^{d\top})(\mathbf{X} - \mathbf{F}^d\mathbf{L}^{d\top})^\top]\}.$$

3. DOUBLE CROSS VALIDATION FOR NUMBER OF FACTORS

Our basic idea to specify the number of factor d_0, is based on the prediction error in a cross-validated manner. Write the model element-wise as

$$(3) \qquad x_{is} = \boldsymbol{f}_i^{0\top} \boldsymbol{l}_s^0 + e_{is}, \quad i = 1, ..., n, \ \ s = 1, 2, ..., p.$$

CV needs to make prediction of x_{is} based on other elements except for itself. Because neither $\boldsymbol{f}_i^0$ nor $\boldsymbol{l}_s^0$ in (3) is observable, we implement the prediction by two-stage fitting: leave-*observation*-out and leave-*variable*-out.

The first stage is to estimate $\boldsymbol{l}_s^0$ from the data by the K-fold CV. Divide the rows of $\mathbf{X}$ into K folds, $1 < K \leq n$. Denote them by $M_1, ..., M_K$. Let $n_k = \#M_k$ be the number of elements in M_k, and $\mathbf{X}_{-M_k}$ be the sub-matrix of $\mathbf{X}$ with rows in M_k being removed. Apply the eigen-decomposition to $\mathbf{X}_{-M_k}$ to obtain corresponding matrices $\widetilde{\mathbf{F}}^{k,d}$ and $\widetilde{\mathbf{L}}^{k,d}$ as in (2). Here, we use the superscript to highlight the fact that the estimator is obtained from the data with the rows in M_k removed, with working number of factors d. Similar to Bai and Ng (2002), we rescale the estimated factor loading matrix and let

$$\widehat{\mathbf{L}}^{k,d} = (\frac{1}{n - n_k}\mathbf{X}_{-M_k}^\top \mathbf{X}_{-M_k})\widetilde{\mathbf{L}}^{k,d} = (\widehat{\boldsymbol{l}}_1^{k,d}, ..., \widehat{\boldsymbol{l}}_p^{k,d}).$$

In the second stage, we replace $\boldsymbol{l}_s^0$ in (3) by $\widehat{\boldsymbol{l}}_s^{k,d}$ and rewrite (3) as a regression model

$$(4) \qquad x_{is} = \boldsymbol{f}_i^{0\top} \widehat{\boldsymbol{l}}_s^{k,d} + e'_{is}, \quad i \in M_k, \ \ s = 1, 2, ..., p,$$

where e'_{is} is the regression error in place of e_{is} due to the replacement. This time, $\widehat{\boldsymbol{l}}_s^{k,d}, s = 1, 2, ..., p$ are known and treated as the 'regressors', but $\boldsymbol{f}_i^0$ is treated as 'regression coefficients' that need to

be estimated. By leaving x_{is} out, we estimate $\boldsymbol{f}_i^0$ by

$$\widehat{\boldsymbol{f}}_{i,s}^{k,d} = \operatorname*{argmin}_{\boldsymbol{f}} \sum_{t \neq s, t=1}^{p} (x_{it} - \boldsymbol{f}^\top \widehat{\boldsymbol{l}}_t^{k,d})^2, \ i \in M_k.$$

Thus we can predict x_{is} by $(\widehat{\boldsymbol{f}}_{i,s}^{k,d})^\top \widehat{\boldsymbol{l}}_s^{k,d}$. The average squared prediction error for $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^\top$ is

$$V_i^{k,d} = \frac{1}{p} \sum_{s=1}^{p} [x_{is} - (\widehat{\boldsymbol{f}}_{i,s}^{k,d})^\top \widehat{\boldsymbol{l}}_s^{k,d}]^2, \quad i \in M_k.$$

The calculation of $V_i^{k,d}$ can be much simplified as shown in Shao (1993), i.e.,

$$V_i^{k,d} = \frac{1}{p} \sum_{s=1}^{p} (1 - w_s^{k,d})^{-2} (x_{is} - (\widehat{\boldsymbol{f}}_i^{k,d})^\top \widehat{\boldsymbol{l}}_s^{k,d})^2,$$

where $\widehat{\boldsymbol{f}}_i^{k,d} = (\widehat{\boldsymbol{L}}^{k,d\top} \widehat{\boldsymbol{L}}^{k,d})^{-1} \widehat{\boldsymbol{L}}^{k,d\top} \boldsymbol{x}_i$ is the conventional least squares estimator for (4) and $w_s^{k,d}$ is the s-th diagonal element of projection matrix $\mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}} = \widehat{\boldsymbol{L}}^{k,d} (\widehat{\boldsymbol{L}}^{k,d\top} \widehat{\boldsymbol{L}}^{k,d})^{-1} \times \widehat{\boldsymbol{L}}^{k,d\top}$. Finally, consider the averaged prediction error over all the elements

$$DCV(d) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k}^{n} V_i^{k,d}.$$

Let $d_{\max}$ be a fixed positive integer that is large enough such that $p > d_{\max} > d_0$. The DCV estimator for the number of factors is given by

$$(5) \qquad \widehat{d}_{DCV} = \operatorname*{argmin}_{0 \leq d \leq d_{\max}} DCV(d).$$

The above approach has similarity with the vanilla row-wise CV (Bro et al., 2008), which also predicts $\boldsymbol{x}_i$ by two steps. Its first step is exactly the same as ours. However, in the second step, row-wise CV uses the full data instead of cross validation method, leading to overfitting and inconsistency of the method. To fix this problem, Eastment and Krzanowski (1982) suggested the element-wise cross validation. Although their methods solve the problem of overfitting, they are costly and involve immense computation.

Our DCV offers a simpler solution. By utilizing CV in both the first stage of leaving "observations" out, and the second regression step of leaving "variables" out, our method guarantees that the prediction of each element does not use any information from itself, and thus the consistency is ensured as shown in the next section. In addition, the implement of K-fold CV can significantly

reduce the computational complexity, especially when the number of rows is large. It is interesting to see that the K-fold CV in the first stage will also help in simplifying the calculation in the second stage. Because of this appealing nature, to further facilitate the calculation, we can transpose $\boldsymbol{X}$ when $n < p$.

4. ASYMPTOTIC CONSISTENCY OF THE ESTIMATION

In this section, we investigate the consistency of our derived estimator. We start with the following assumptions.

ASSUMPTION 1   *The eigenvalues of $\boldsymbol{F}^{0\top}\boldsymbol{F}^0/n$ are bounded away from zero and infinity, and that $\mathbb{E}\|\boldsymbol{f}_i^0\|^4 < \infty$ for $i = 1, ..., n$.*

ASSUMPTION 2   *The eigenvalues of $\boldsymbol{L}^{0\top}\boldsymbol{L}^0/p$ are bounded away from zero and infinity, and that $\mathbb{E}\|\boldsymbol{l}_i^0\|^4 < \infty$ for $i = 1, ..., p$.*

ASSUMPTION 3   *There exists a constant $M < \infty$ such that*

1. $\mathbb{E}e_{it} = 0$ *and* $\mathbb{E}e_{it}^4 \leq M$ *for* $1 \leq i \leq n$ *and* $1 \leq t \leq p$;

2. $\gamma_p(i,j) = \mathbb{E}(\boldsymbol{e}_i^\top \boldsymbol{e}_j/p) \leq M$ *and* $\frac{1}{n}\sum\limits_{j=1}^{n}\gamma_p^2(i,j) \leq M$ *for all* i;

3. $\mathbb{E}(\boldsymbol{e}_{is}\boldsymbol{e}_{it}) = \tau_{st,i}$ *with* $\tau_{st,i} \leq |\tau_{st}|$ *for some* τ_{st} *and for all* i; *in addition,* $\frac{1}{p}\sum\limits_{s=1}^{p}\sum\limits_{t=1}^{p}\tau_{st} \leq M$;

4. $\mathbb{E}(\boldsymbol{e}_{is}\boldsymbol{e}_{jt}) = \tau_{ij,st}$ *and* $\frac{1}{np}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\sum\limits_{s=1}^{p}\sum\limits_{t=1}^{p}\tau_{ij,st} \leq M$;

5. $\frac{1}{n}\mathbb{E}|\sum\limits_{j=1}^{n}[e_{jt}e_{js} - \mathbb{E}(e_{jt}e_{js})]|^2 \leq M$ *for all* (s,t);

6. $\|\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}\boldsymbol{f}_i^0 e_{is}\|^2 \leq M$ *for* $1 \leq s \leq p$;

7. $\|\frac{1}{\sqrt{p}}\sum\limits_{s=1}^{p}\boldsymbol{l}_s^0 e_{is}\|^2 \leq M$ *for* $1 \leq i \leq n$.

Assumptions 1–3 are commonly used in the analysis of factor models, for example, in Bai and Ng (2002) and Li et al. (2017). Assumptions 1 and 2 together ensure that each factor plays a nontrivial role in contributing to the variation of $\boldsymbol{X}$. Unlike the strict factor model that assumes all entries of $\boldsymbol{E}$ are I.I.D., Assumption 3 allows weak dependency for elements of $\boldsymbol{E}$, making our methods applicable to the approximate factor models. These assumptions also indicate that all the eigenvalues of the covariance matrix of common components would dominate the eigenvalues

corresponding to idiosyncratic components, which is crucial to the identifiability for the approximate factor models.

ASSUMPTION 4 $\sup_{k} n_k = o(n^{1/3})$ and $K \sup_{k} n_k/n < \infty$, where $n_k = \#M_k$ is the number of elements in M_k.

Assumption 4 is a weak condition on K-fold CV in the first stage. The number of elements in each fold can either be fixed or tend to infinity. In particular, this assumption includes the leave-one-out CV.

ASSUMPTION 5.a *The idiosyncratic components satisfies*

1. $\mathbb{E}[e_{it}^2 \mid \boldsymbol{F}^0_{-M_k} \boldsymbol{L}^{0\top}] = \sigma^2$ for $i \in M_k$, where $\boldsymbol{F}^0_{-M_k}$ is a submatrix of $\boldsymbol{F}^0$ with rows in M_k being removed;
2. $\lambda_1(\boldsymbol{E}^\top \boldsymbol{E}/n) < 2\sigma^2$.

ASSUMPTION 5.b *Let $p/n \to \rho \in (0, \infty)$. There exists a $p \times p$ positive definite matrix $\boldsymbol{\Sigma}_p$ such that $\boldsymbol{e}_i = \boldsymbol{\Sigma}_p^{1/2} \boldsymbol{y}_i$, where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{ip})^\top$,*

1. $\mathbb{E}y_{it} = 0$, $\mathbb{E}y_{it}^4 < \infty$, and y_{it} are I.I.D. for $1 \leq i \leq n, 1 \leq t \leq p$;
2. $\boldsymbol{y}_i$ is independent of $\boldsymbol{F}^0_{-M_k}$;
3. The spectral distribution of $\boldsymbol{\Sigma}_p$ is convergent, i.e., there exists a distribution function H such that $\frac{1}{p} \sum_{s=1}^{p} \mathbb{1}(\lambda_s(\boldsymbol{\Sigma}_p) \leq t) \to H(t)$ as $p \to \infty$;
4. $2\text{diag}(\boldsymbol{\Sigma}_p) - \boldsymbol{\Sigma}_p$ is a positive definite matrix, where $\text{diag}(\boldsymbol{\Sigma}_p)$ is a diagonal matrix that has the same diagonal elements as $\boldsymbol{\Sigma}_p$.

Assumptions 5.a and 5.b are two technical assumptions, either of which ensures the viability of the proposed method. They also allow weak dependence among idiosyncratic components as well as the observations, which generalize the I.I.D. assumption required in the strict factor model. In comparison, Bai and Ng (2002) needs weaker assumptions on the dependence than ours, possibly due to the relatively large penalty in their methods. Of course, this large penalty is at the cost of selecting inappropriately fewer factors as shown in our simulation study.

THEOREM 1 *Suppose Assumptions 1–4 hold. If either Assumption 5.a or 5.b holds, then the DCV*

estimator (5) *is consistent as both n and p tend to infinity, i.e.,*

$$\lim_{n,p\to\infty} Prob(\widehat{d}_{DCV} = d_0) = 1.$$

REMARK 1   Recent empirical findings suggest that the factor structure of economic data may change with sample size and dimensionality of variable (Onatski, 2010; Jurado et al., 2015; Li et al., 2017). In fact, Theorem 1 can be easily extended to the case that d_0 varies slowly with both n and p. DCV is thus applicable to practical data with changing structure of factorization.

5. SIMULATION STUDY

In this section, we conduct simulation studies to compare the finite sample performances of the DCV with the other methods, including the panel criterion IC_1 (Bai and Ng, 2002) and *Ladle* (Luo and Li, 2016). As shown in Luo and Li (2016), *Ladle* estimator outperforms all the other eigen-structrue based methods in determining the number of factors, and is thus used as a representative of those methods. Similarly, for the IC-based methods, Bai and Ng (2002) showed that their criterion outperforms other information criteria such as AIC and BIC. We do not consider other cross-validation methods as they are either inconsistent or suffer from excessive computational burden. For our DCV, we consider both leave-one-out CV (DCV_1) and 10-fold CV (DCV_{10}) in the first stage. Matlab codes for the calculations are available at https://github.com/XianliZeng/Double-Cross-Validation.

We simulate data from a factor model with $d_0 = 5$:

$$(6) \qquad x_{is} = \sum_{j=1}^{5} f_{ij}l_{sj} + \sqrt{\theta}e_{is}; \quad f_{ij}, l_{sj} \sim N(0,1) \ \ i = 1, ..., n, \ \ s = 1, 2, ..., p.$$

For the idiosyncratic errors e_{is}, the following settings are considered:

(E1) independent Gaussian: $e_{is} \sim N(0,1)$;

(E2) independent t-distributed: $e_{is} \sim t_3$;

(E3) heteroskedastic: $e_{is} \sim N(0, \delta_s)$, $\delta_s = 1$ if s is odd, and 2 if s is even;

(E4) serial correlated: $e_{is} = 0.3e_{i,s-1} + \nu_{is}$, $\nu_{is} \sim N(0,1)$;

(E5) cross-sectional correlated: $e_{is} = \sum_{j=-10}^{10} 0.15^j \nu_{i-j,s}$, $\nu_{is} \sim N(0,1)$.

In all simulations, we select d from $\{1, 2, ..., d_{\max}\}$ where $d_{\max} = 8$ as in Bai and Ng (2002). We consider the combinations of n and p with $n = \{40, 160, 640\}$ and $p = \{30, 90, 270\}$. To see
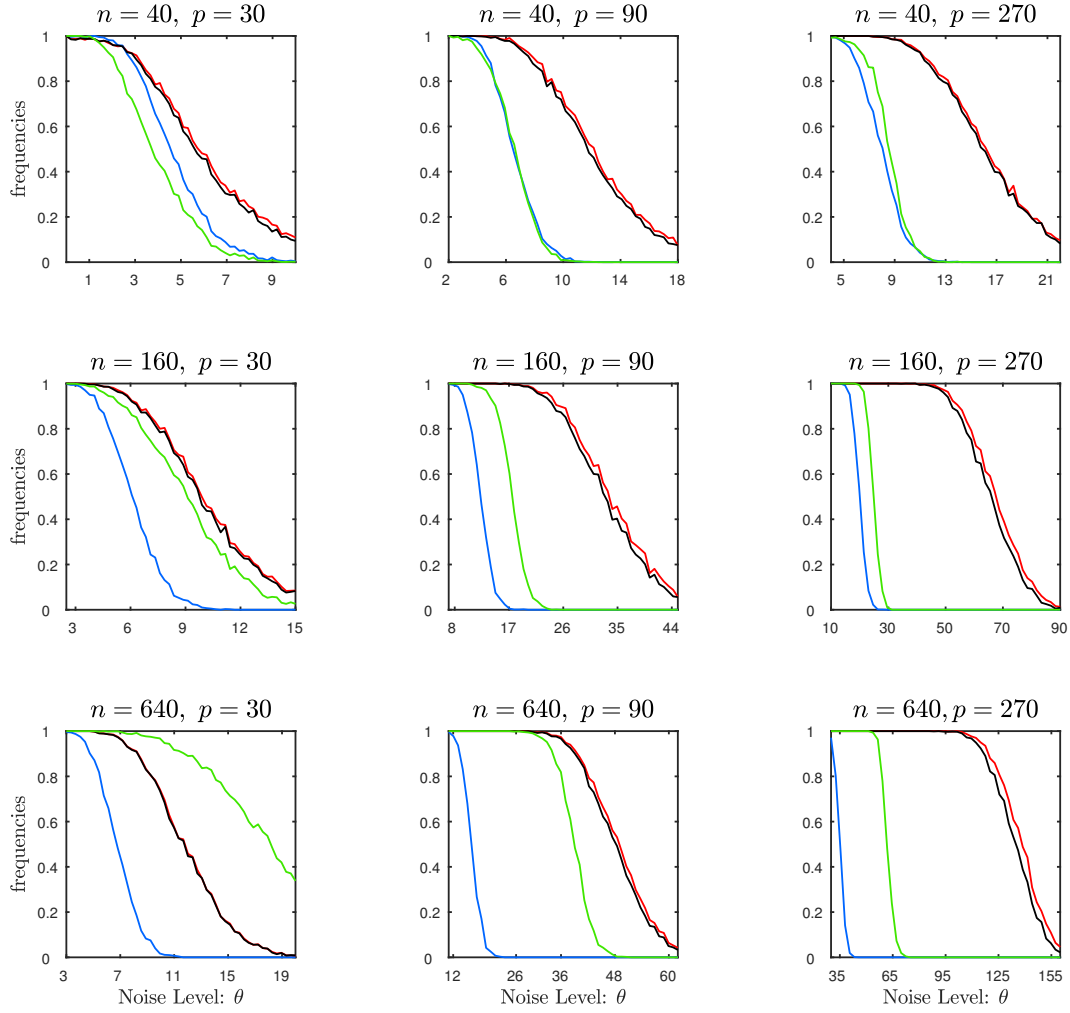
FIGURE 1.— The relative frequencies of correctly selecting the number of factors for model (6) with independent Gaussian idiosyncratic errors (E1). In each panel, we use red line for DCV, black line for DCV_{10}, blue line for IC_1 and green line for *Ladle*.

the effect of the signal-to-noise ratio on the methods, θ varies in a range of intervals depending on different settings of idiosyncratic errors and different n and p. These ranges are made clear in each panel of the figures below. All comparisons were evaluated based on 1000 replications. Figures 1 to 5 summarize the simulation results, where the curves represent the frequencies of selecting the right number of factors at different noise level θ.

Generally, the simulation results suggest that the performance of all the methods becomes
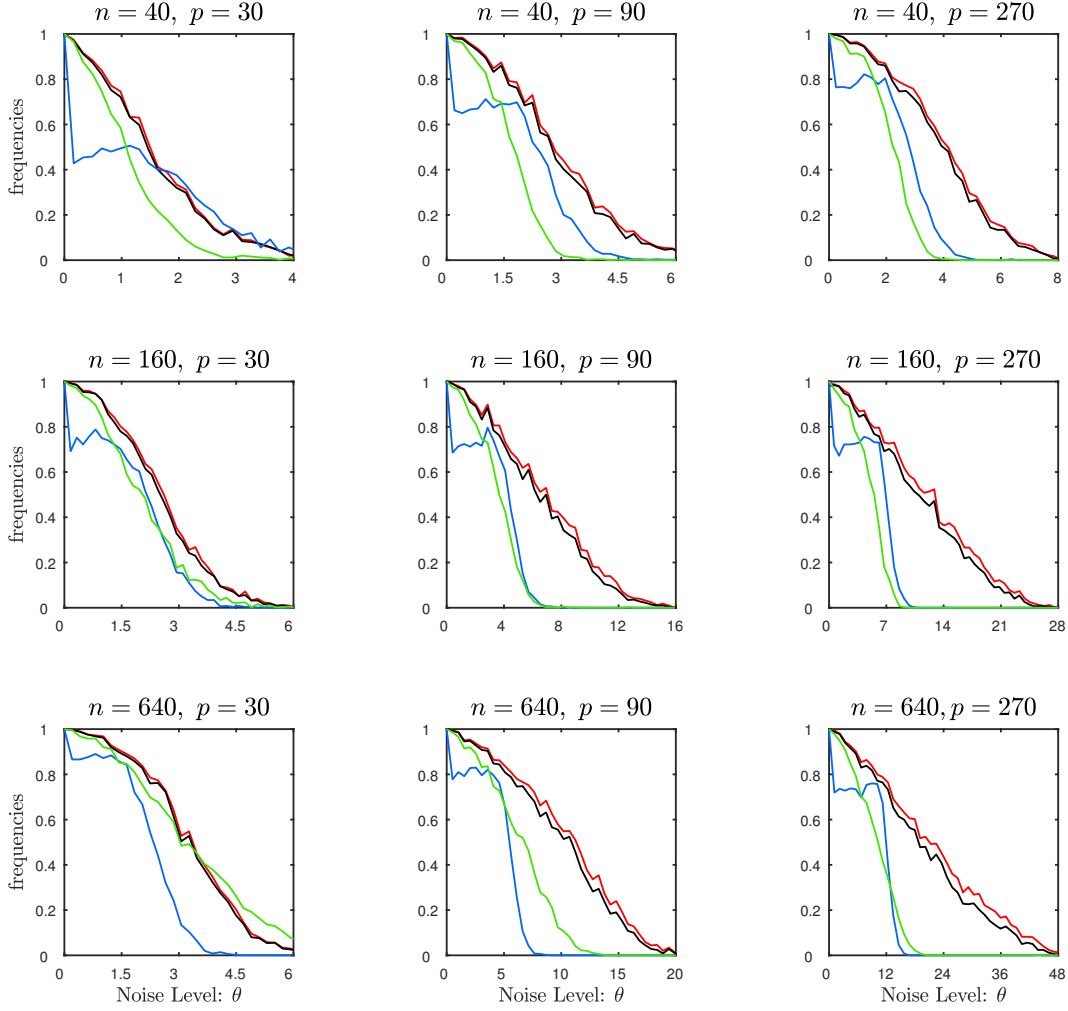
FIGURE 2.— The relative frequencies of correctly selecting the number of factors for model (6) with independent t-distributed idiosyncratic errors (E2). In each panel, we use red line for DCV, black line for DCV_{10}, blue line for IC_1 and green line for $Ladle$.

better when n and p get larger, which lends support to the asymptotic consistency of the methods. $Ladle$ usually shows better performance than IC_1, possibly due to its utilization of both eigenvalues and eigenvectors. Notably, DCV_1 and DCV_{10} stand out and perform the best in most scenarios. They possess a much higher frequency of correct selection when θ is large, implying their ability to select the right number of factors even when the factors' contribution to the variation of the variables, i.e., signal-to-noise ratio, is low, which is particularly the situation in economic data.

FIGURE 3.— The relative frequencies of correctly selecting the number of factors for model (6) with heteroskedastic idiosyncratic errors (E3). In each panel, we use red line for DCV, black line for DCV_{10}, blue line for IC_1 and green line for *Ladle*.

Moreover, DCV_1 and DCV_{10} present almost identical performance for all the models, implying that the performance of the DCV is not sensitive to the choice of K.

For different settings of the idiosyncratic errors, we have some additional observations. Figure 2 shows that IC_1 is not so robust to the data with heavy-tail especially when the size of the data is small. Figure 3 implies that IC_1 is also affected adversely by the heteroskedastic errors. Note that both heavy tail and heteroskedasticity are stylized facts in financial data. In contrast, for these two
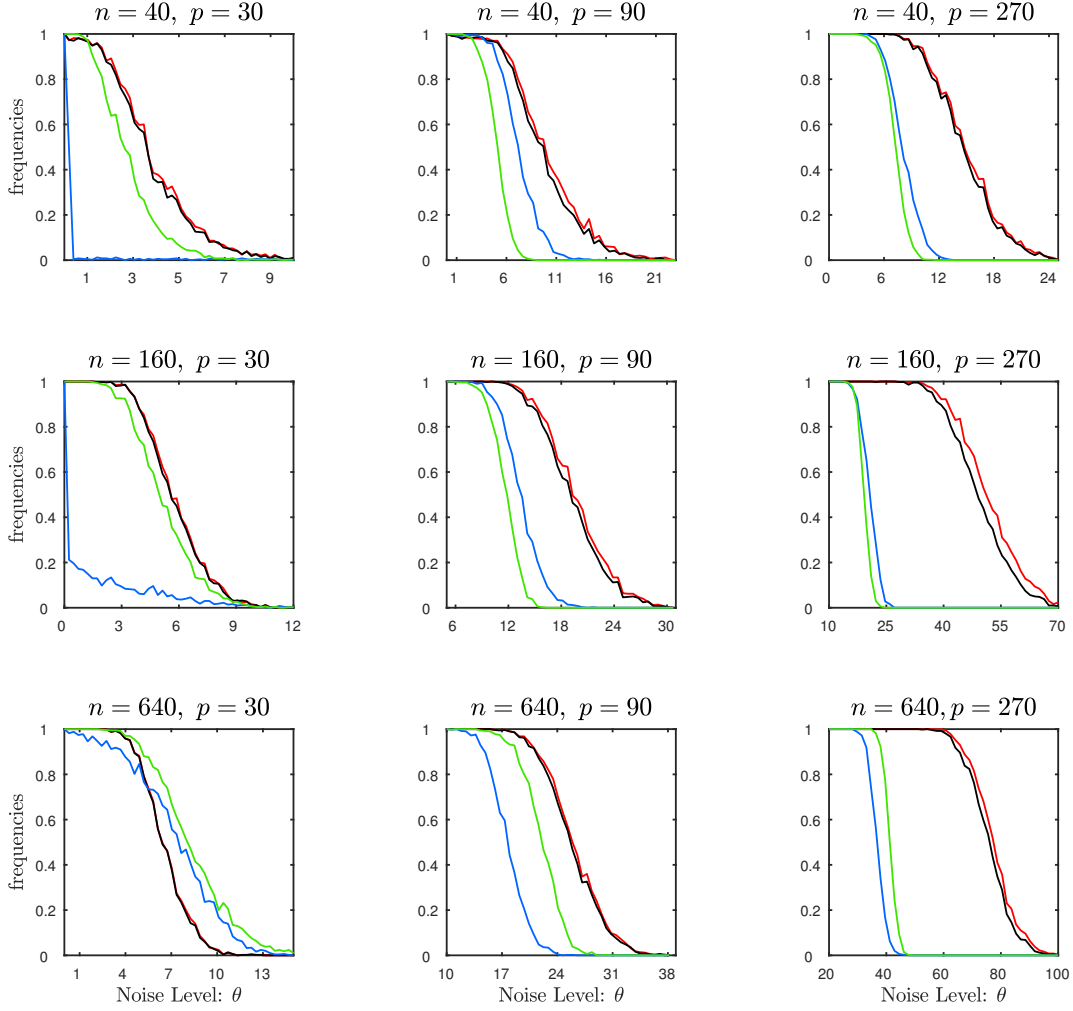
FIGURE 4.— The relative frequencies of correctly selecting the number of factors for model (6) with serial correlated idiosyncratic errors (E4). In each panel, we use red line for DCV, black line for DCV_{10}, blue line for IC_1 and green line for *Ladle*.

types of data, DCV_1 and DCV_{10} still demonstrate very stable and efficient ability in selecting the number of factors. In Figures 4 and 5, where the idiosyncratic errors have serial or cross-sectional correlation, DCV has inferior performance than IC_1 and *Ladle* when the variance of idiosyncratic components is small and both n and p are small. However, DCVs quickly gains advantage when either n or p increases.
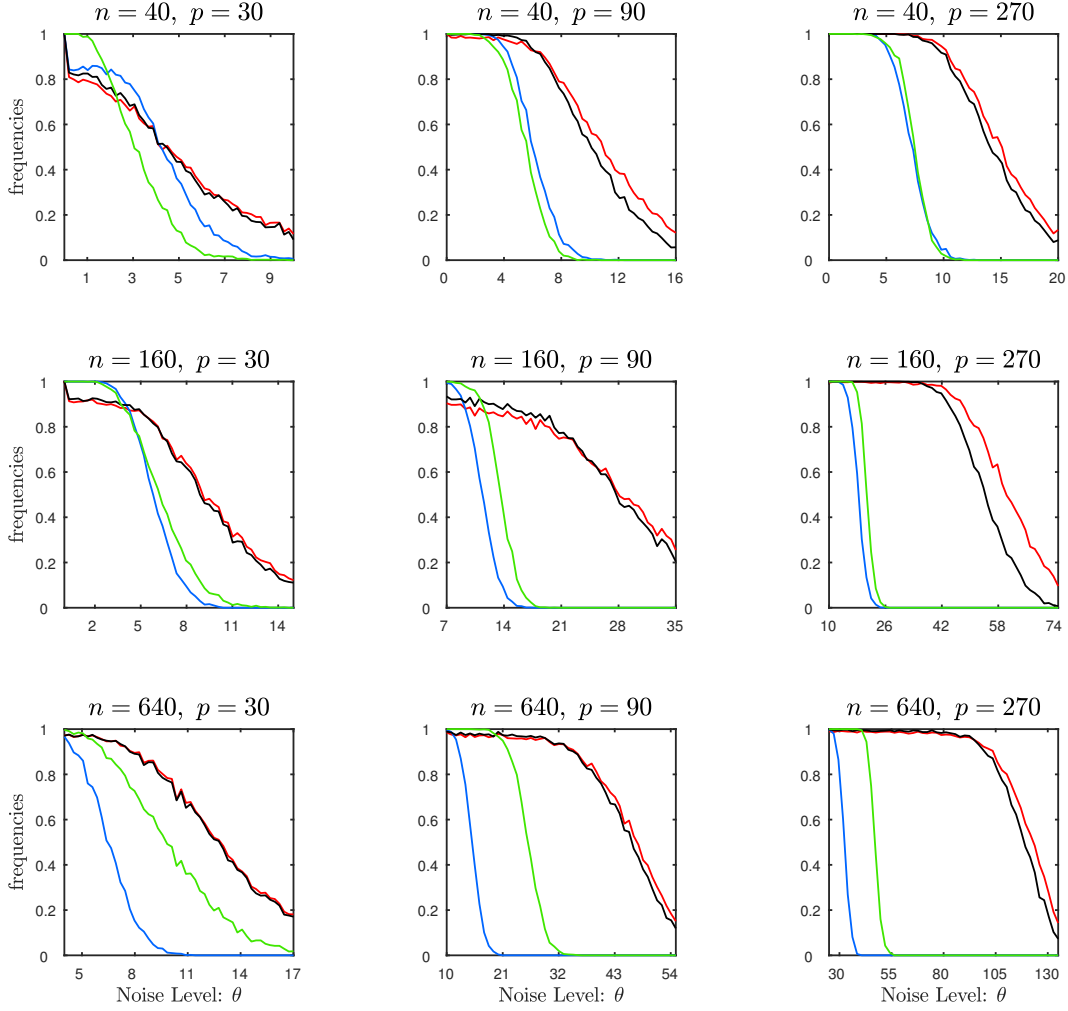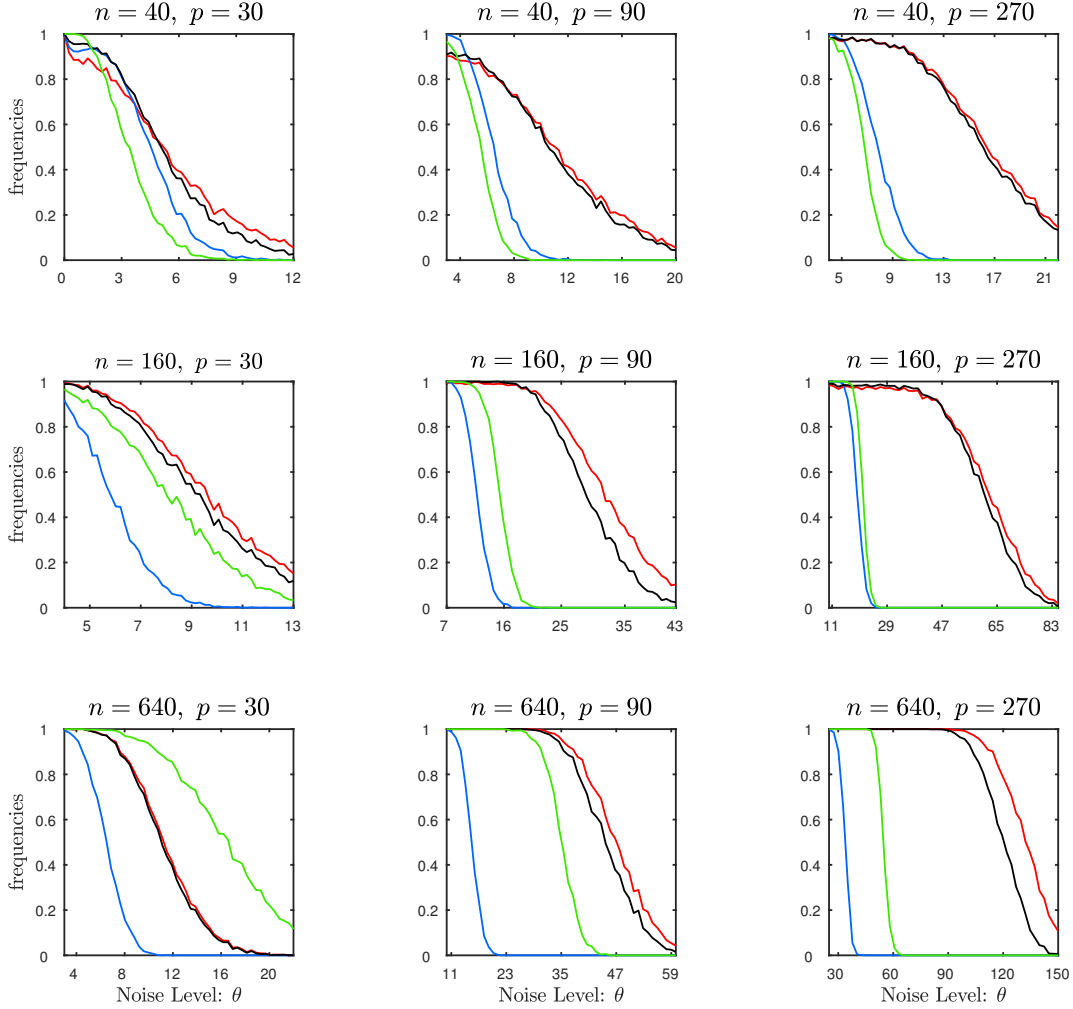
FIGURE 5.— The relative frequencies of correctly selecting the number of factors for model (6) with cross-sectional correlated idiosyncratic errors (E5). In each panel, we use red line for DCV, black line for DCV_{10}, blue line for IC_1 and green line for $Ladle$.

6. AN EMPIRICAL APPLICATION: FAMA-FRENCH THREE-FACTOR MODEL

In asset pricing and portfolio management, the so-called three-factor model of Fama and French (1993) is widely applied to describe the stock returns. It shows that the excess return of a stock or a portfolio $(R_{it} - r_{ft})$ can be satisfactorily explained by three common factors: (1) excess return of the market portfolio $(R_{Mt} - r_{ft})$, (2) the outperformance of small-cap companies versus big-cap companies (SMB), and (3) the outperformance of companies with high book to market ratio versus

Validation of three-factor model based on value weighted return of 25 portfolios
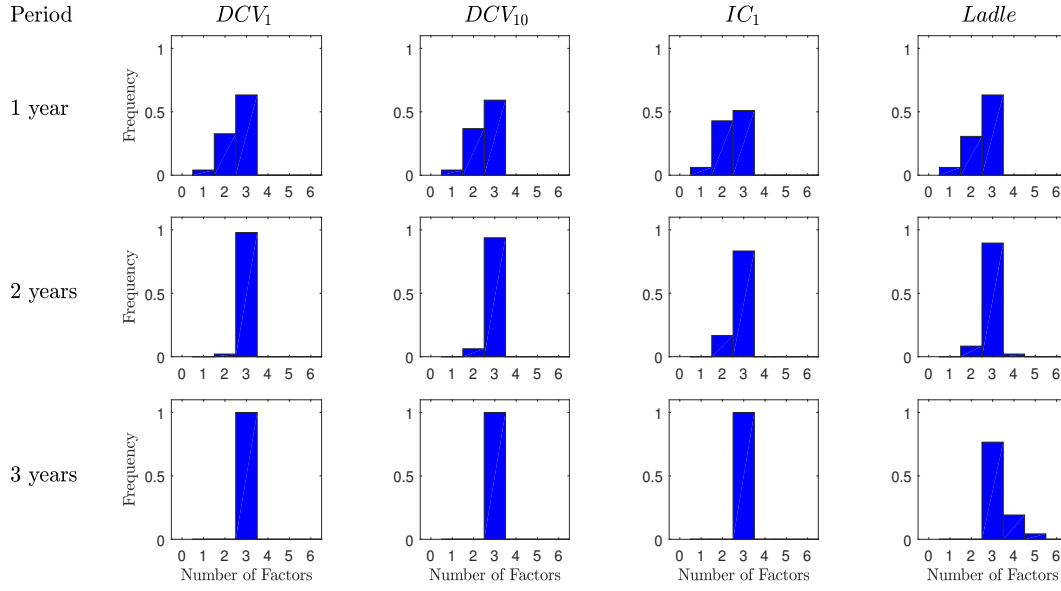
Validation of three-factor model based on value weighted return of 100 portfolios

FIGURE 6.— The frequencies of selected numbers of factors for the value weighted returns of 25 and 100 portfolios. Panels in each column represents one method; panels in the same rows are based on the same data in fixed period of years.

those with small book to market ratio (HML), i.e.,

$$R_{it} - r_{ft} = \beta_1(R_{Mt} - r_{ft}) + \beta_2 SMB_t + \beta_3 HML_t + \epsilon_{it},$$

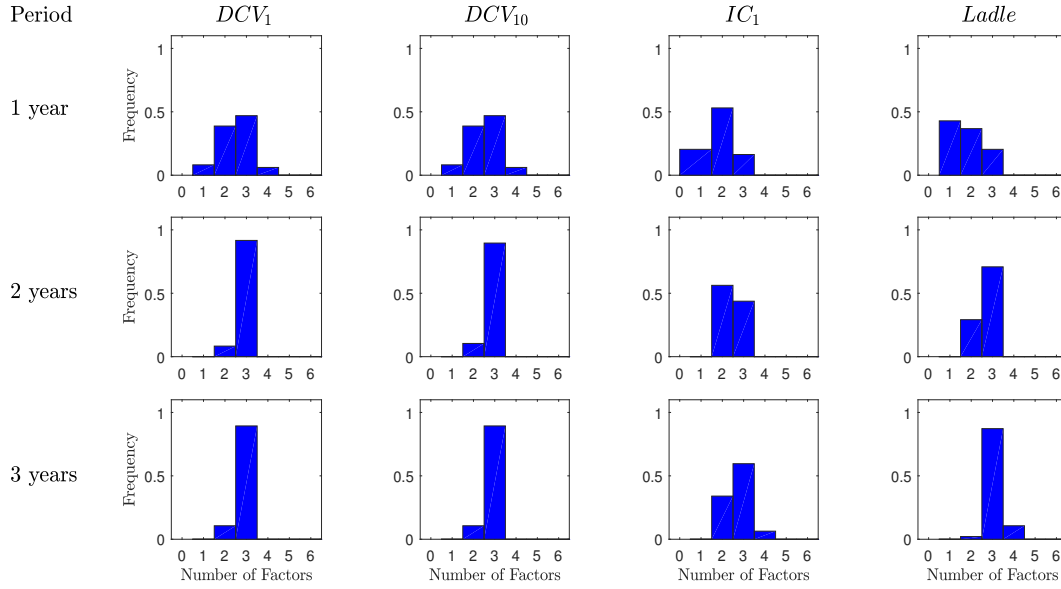where r_{ft} is the risk-free return. Next, we use the data provided by Professor Kenneth R. French[1] to mutually validate our methods and the three-factor model. Stocks are divided into 6, 25 or 100 groups according to company's capitalization (market equity) and Book-to-Market ratio, and portfolios are constructed with value weighted daily returns or equally weighted daily returns in each group. As those portfolios were compiled based on company's data at the end of June of every year, we set the first of July of every year as a starting date of a period. It is understandable that the portfolios change from year to year, thus we only consider periods of 1 to 3 years in our calculation.

Figure 6 presents the frequencies of the numbers of factors detected by the DCV_1, DCV_{10}, IC_1 and $Ladle$ in two of the data sets with $d_{\max}$ set at 15. In most situations, all the approaches identify three common factors for the returns of portfolios, which are in line with the *three*-factor model of Fama and French (1993). In general, DCV selects three factors most often amongst the approaches, showing its superior performance in practice.

7. CONCLUDING REMARKS

In the literature, both CV and IC based approaches were popularly used for determining the complexity of a model, such as the number of factors. As noticed in Bro et al. (2008), most CV approaches for factor models are not consistent due to their insufficient validation, while existing remedies for the consistency is very computationally expensive. In contrast, ICs are consistent and easy to implement, but they depend on predetermined penalty functions (Bai and Ng, 2002). Simulation results not reported here show that the penalty function in Bai and Ng (2002) could be modified to be more adaptive to the data. In addition, ICs are unstable when the signal-to-noise ratio is relatively low (Onatski, 2010).

By validating the model twice, the proposed DCV method does not only ensure the consistency under mild conditions, but is also easy to implement. Because DCV is based on prediction error, it automatically selects the number of factors that balances the model complexity and stability. Our simulation studies also demonstrate its superior efficiency over the existing methods most of the time. The only exception is when there is strong serial dependence in the idiosyncratic errors. However, this

---

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

deficiency disappears when the sample size and dimension of the variables increase. The advantages of our method are more pronounced for data with relative large variation, heteroscedasticity or heavy tails in the idiosyncratic errors. The method is thus particularly relevant for financial data.

REFERENCES

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, Z. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442.

Bao, Z., Pan, G., and Zhou, W. (2015). Universality for the largest eigenvalue of sample covariance matrices with general population. *The Annals of Statistics*, 43(1):382–421.

Bro, R., Kjeldahl, K., Smilde, A. K., and Kiers, H. A. L. (2008). Cross-validation of component models: a critical look at current methods. *Analytical and bioanalytical chemistry*, 390(5):1241–1251.

Cragg, J. G. and Donald, S. G. ((1997)). Inferring the rank of a matrix. *Journal of Econometrics*, 76:223–250.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

Eastment, H. T. and Krzanowski, W. J. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, 24(1):73–77.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4):540–554.

Fujikoshi, Y. (1977). Asymptotic expansions for the distributions of some multivariate tests. In *Krishnaiah, P. R., ed. Multivariate Analysis*, volume IV, pages 55–71, Amsterdam: North-Holland.

Gregory, A. W. and Head, A. C. (1999). Common and country-specific fluctuations in productivity, investment, and the current account. *Journal of Monetary Economics*, 44(3):423–451.

Horn, R. A. and Johnson, C. R. (1990). *Matrix analysis*. Cambridge university press.

Jurado, K., Ludvigson, S. C., and Ng, S. (2015). Measuring uncertainty. *American Economic Review*, 105(3):1177–1216.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264.

Lewbel, A. (1991). The rank of demand systems: theory and nonparametric estimation. *Econometrica*, 59:711–730.

Li, H., Li, Q., and Shi, Y. (2017). Determining the number of factors when the number of factors can increase with

sample size. *Journal of Econometrics*, 197(1):76–86.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103(4):875–887.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016.

Ross, S. A. (1976). The arbitrage theory of capital asset pricingheory of capital asset pricing. *Journal of Finance*, 13:341–360.

Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, 89(425):141–148.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494.

Shao, J. (2003.). *Mathematical statistics. 2nd ed.* New York: Springer.

Stock, J. H. and Watson, M. W. (1998). Diffusion indexes. *NBER working paper*.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):968–979.

Yu, Y., Wang, T., and Samworth, R. J. (2014). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

APPENDIX A: PROOFS FOR THEOREMS

In this section, we provide detailed proofs for our theoretical results, mainly based on the advanced techniques derived in Bai and Ng (2002) and RMT (Bai and Zhou, 2008; Ledoit and Péché, 2011; Bao et al., 2015). We first introduce several notations that repeatedly appear hereafter. For matrix $\boldsymbol{A} = (a_{ij})^{n \times p}$, $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \cdots, a_{ip})^{\top}$ is the transpose of the i-th row, and $\boldsymbol{A}_{-M_k}$ is the sub-matrix of $\boldsymbol{A}$ with rows in M_i being removed, and $\lambda_t(\boldsymbol{A})$ is the t-th largest eigenvalue of $\boldsymbol{A}$, and $\|\boldsymbol{A}\|$ is the Frobenius norm of $\boldsymbol{A}$. Let $\widehat{\lambda}_t, t = 1, 2, ..., p$ be the eigenvalues of $\boldsymbol{X}^{\top}\boldsymbol{X}$ with decreasing order and $\widehat{\phi}_t, t = 1, 2, ..., p$ be the corresponding eigenvectors. Let $(\lambda_t, \phi_t, t = 1, ..., p)$, $(\widehat{\lambda}_t^k, \widehat{\phi}_t^k, t = 1, ..., p)$ and $(\lambda_t^k, \phi_t^k, t = 1, ..., p)$ be corresponding quantities for $\boldsymbol{L}^0\boldsymbol{F}^{0\top}\boldsymbol{F}^0\boldsymbol{L}^{0\top}$, $\boldsymbol{X}_{-M_k}^{\top}\boldsymbol{X}_{-M_k}$ and $\boldsymbol{L}^0\boldsymbol{F}_{-M_k}^{0\top}\boldsymbol{F}_{-M_k}^0\boldsymbol{L}^{0\top}$, respectively. We use M_{np} to denote $\max(n, p)$ and m_{np} to denote $\min(n, p)$. In addition, we use C and C' to denote constants that may vary in different places throughout the paper.

To make the proofs easy to read, we give the proof to our main result in the first subsection, and put supporting facts and their proofs in the second subsection of this Appendix.

A.1. *Proof of Theorem 1*

Theorem 1 follows if we can show that

$$\lim_{n,p\to\infty} \text{Prob}(DCV(d) < DCV(d_0)) = 0 \text{ for } d \neq d_0 \text{ and } 1 \leq d \leq d_{\max}.$$

Let $w_s^{k,d}$ be the sth diagonal element of $\mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}} = \widehat{\boldsymbol{L}}^{k,d}(\widehat{\boldsymbol{L}}^{k,d\top}\widehat{\boldsymbol{L}}^{k,d})^{-1}\widehat{\boldsymbol{L}}^{k,d\top}$, $W_i(d,\boldsymbol{L}) = \frac{1}{p}\boldsymbol{x}_i^\top(\boldsymbol{I}_d - \mathcal{P}_{\boldsymbol{L}})\boldsymbol{x}_i$ and $\boldsymbol{H}^{k,d} = \frac{1}{(n-n_k)p}\boldsymbol{F}_{-M_k}^{0\top}\boldsymbol{X}_{-M_k}^\top\widetilde{\boldsymbol{L}}^{k,d}$. For $1 \leq d < d_0$, according to Shao (1993),

$$
\begin{aligned}
DCV(d) &= \frac{1}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\sum_{s=1}^{p}(1-w_s^{k,d})^{-2}(x_{is} - \widehat{\boldsymbol{f}}_i^{k,d\top}\widehat{\boldsymbol{l}}_s^{k,d})^2 \\
&= \frac{1}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\sum_{s=1}^{p}(1+2w_s^{k,d}+O(w_s^{k,d^2}))(x_{is} - \widehat{\boldsymbol{f}}_i^{k,d\top}\widehat{\boldsymbol{l}}_s^{k,d})^2 \\
&= \frac{1}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\boldsymbol{x}_i^\top(\boldsymbol{I}_d - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}})\boldsymbol{x}_i + o_p(1) \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d,\widehat{\boldsymbol{L}}^{k,d}) + o_p(1).
\end{aligned}
$$

Then,

$$
\begin{aligned}
DCV(d) - DCV(d_0) &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d,\widehat{\boldsymbol{L}}^{k,d}) - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d_0,\widehat{\boldsymbol{L}}^{k,d_0}) + o_p(1) \\
&= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d,\widehat{\boldsymbol{L}}^{k,d}) - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d,\boldsymbol{L}^0\boldsymbol{H}^{k,d}) \\
&\quad + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d,\boldsymbol{L}^0\boldsymbol{H}^{k,d}) - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d_0,\boldsymbol{L}^0) \\
&\quad + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d_0,\boldsymbol{L}^0) - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d_0,\boldsymbol{L}^0\boldsymbol{H}^{k,d_0}) \\
&\quad + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d_0,\boldsymbol{L}^0\boldsymbol{H}^{k,d_0}) - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}W_i(d_0,\widehat{\boldsymbol{L}}^{k,d_0}) + o_p(1) \\
&> c_d + o_p(1).
\end{aligned}
$$

Here, the last inequality follows from Lemma 3, Lemma 4 and the fact that $\mathcal{P}_{\boldsymbol{L}^0} = \mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d_0}}$. Consequently,

$$\lim_{n,p\to\infty} \text{Prob}(DCV(d) < DCV(d_0)) = 0.$$

Let $\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} = \text{diag}(w_1^{k,d}, w_2^{k,d}, ..., w_p^{k,d})$ and $\text{Cov}(\boldsymbol{e}_i) = \boldsymbol{\Sigma}_{p,i}$. For $d_0 < d \leq d_{\max}$, according to Lemma 7,

$$
\begin{aligned}
&DCV(d) \\
&= \frac{1}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\sum_{s=1}^{p}(1-w_s^{k,d})^{-2}(x_{is} - \widehat{\boldsymbol{f}}_i^{k,d\top}\widehat{\boldsymbol{l}}_s^{k,d})^2
\end{aligned}
$$

$$
= \frac{1}{np} \sum_{k=1}^{K} \sum_{i \in M_k} \sum_{s=1}^{p} (1 + 2w_s^{k,d} + O(w_s^{k,d^2}))(x_{is} - \widehat{\boldsymbol{f}}_i^{k,d\top} \widehat{\boldsymbol{l}}_s^{k,d})^2
$$

$$
= \frac{1}{np} \sum_{k=1}^{K} \sum_{i \in M_k} \boldsymbol{x}_i^\top (\boldsymbol{I}_d - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}}) \boldsymbol{x}_i + \frac{2}{np} \sum_{k=1}^{K} \sum_{i \in M_k} \sum_{s=1}^{p} w_s^{k,d} (x_{is} - \widehat{\boldsymbol{f}}_i^{k,d\top} \widehat{\boldsymbol{l}}_s^{k,d})^2 + o_p(\frac{1}{p})
$$

$$
= \frac{1}{np} \sum_{k=1}^{K} \sum_{i \in M_k} \boldsymbol{x}_i^\top (\boldsymbol{I}_d - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}}) \boldsymbol{x}_i + \frac{2}{np} \sum_{k=1}^{K} \sum_{i \in M_k} \text{Tr}(\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} \boldsymbol{\Sigma}_{p,i}) + o_p(\frac{1}{p}).
$$

Therefore,

$$
DCV(d) - DCV(d_0)
$$

$$
= \frac{1}{np} \sum_{k=1}^{K} \sum_{i \in M_k} \boldsymbol{x}_i^\top (\mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d_0}} - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}}) \boldsymbol{x}_i + \frac{2}{np} \sum_{k=1}^{K} \sum_{i \in M_k} \text{Tr}((\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} - \mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d_0}}) \boldsymbol{\Sigma}_{p,i}) + o_p(\frac{1}{p})
$$

$$
= -\frac{1}{np} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 + \frac{2}{np} \sum_{k=1}^{K} \sum_{i \in M_k} \text{Tr}((\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} - \mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d_0}}) \boldsymbol{\Sigma}_{p,i}) + o_p(\frac{1}{p}).
$$

According to Lemma 14,

$$
\frac{1}{n} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 < \frac{2}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \text{Tr}((\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} - \mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d_0}}) \boldsymbol{\Sigma}_{p,i}) + o_p(1).
$$

It follows that,

$$
\lim_{n,p \to \infty} \text{Prob}(DCV(d) < DCV(d_0)) = 0.
$$

We thus complete the proof of Theorem 1.

A.2. *Supporting Lemmas and their Proofs*

LEMMA 1 *Suppose Assumptions 1–4 hold. Let $\widehat{\boldsymbol{l}}_s^{k,d}$ and $\boldsymbol{l}_s^0$ be the transpose of the s-th row of $\widehat{\boldsymbol{L}}^{k,d}$ and $\boldsymbol{L}^0$ respectovely. For any k and d, there exist a $(d_0 \times d)$ matrix $\boldsymbol{H}^{k,d}$ with $\text{rank}(\boldsymbol{H}^{k,d}) = \min(d_0, d)$ such that,*

$$
\|\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top} \boldsymbol{l}_s^0\|^2 = O_p(m_{np}^{-1}), \qquad \text{for each k and s.}
$$

PROOF: Let $\boldsymbol{H}^{k,d} = \frac{1}{(n-n_k)p} \boldsymbol{F}_{-M_k}^{0\top} \boldsymbol{X}_{-M_k}^\top \widetilde{\boldsymbol{L}}^{k,d}$, then,

$$
\widehat{\boldsymbol{L}}^{k,d} - \boldsymbol{L}^0 \boldsymbol{H}^{k,d} = \frac{1}{(n-n_k)p} (\boldsymbol{X}_{-M_k}^\top - \boldsymbol{L}^0 \boldsymbol{F}_{-M_k}^{0\top}) \boldsymbol{X}_{-M_k} \widetilde{\boldsymbol{L}}^{k,d}
$$

$$
= \frac{1}{(n-n_k)p} \boldsymbol{E}_{-M_k}^\top \boldsymbol{X}_{-M_k} \widetilde{\boldsymbol{L}}^{k,d}
$$

$$
= \frac{1}{(n-n_k)p} \boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k} \widetilde{\boldsymbol{L}}^{k,d} + \frac{1}{(n-n_k)p} \boldsymbol{E}_{-M_k}^\top \boldsymbol{F}_{-M_k}^0 \boldsymbol{L}^{0\top} \widetilde{\boldsymbol{L}}^{k,d}.
$$

Thus,

$$
\|\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top} \boldsymbol{l}_s^0\|^2 = \left\| \frac{1}{p} \sum_{t=1}^{p} \widetilde{\boldsymbol{l}}_t^{k,d} \left[\frac{1}{n-n_k} \sum_{i \notin M_k} e_{it} e_{is} \right] + \frac{1}{p} \sum_{t=1}^{p} \widetilde{\boldsymbol{l}}_t^{k,d} \boldsymbol{l}_t^{0\top} \frac{1}{n-n_k} \sum_{i \notin M_k} \boldsymbol{f}_i^0 e_{is} \right\|^2
$$

$$
\leq \quad 3 \left\| \frac{1}{p} \sum_{t=1}^{p} \widetilde{\boldsymbol{l}}_t^{k,d} \left(\frac{1}{n-n_k} \sum_{i \notin M_k} [e_{it} e_{is} - \mathbb{E}(e_{it} e_{is})] \right) \right\|^2
$$

$$
+ 3 \left\| \frac{1}{p} \sum_{t=1}^{p} \widetilde{\boldsymbol{l}}_t^{k,d} \frac{1}{n-n_k} \sum_{i \notin M_k} \mathbb{E}(e_{it} e_{is}) \right\|^2
$$

$$
+ 3 \left\| \frac{1}{p} \sum_{t=1}^{p} \widetilde{\boldsymbol{l}}_t^{k,d} \boldsymbol{l}_t^{0\top} \frac{1}{n-n_k} \sum_{i \notin M_k} \boldsymbol{f}_i^0 e_{is} \right\|^2
$$

$$
= \quad 3[a(k,t) + b(k,t) + c(k,t)].
$$

Next, we investigate the three terms separately as follows.

$$
\begin{aligned}
a(k,t) &\leq \quad \frac{1}{p^2} \sum_{t=1}^{p} \|\widetilde{\boldsymbol{l}}_t^{k,d}\|^2 \sum_{t=1}^{p} \left(\frac{1}{n-n_k} \sum_{i \notin M_k} [e_{it} e_{is} - \mathbb{E}(e_{it} e_{is})] \right)^2 \\
&= \quad \frac{d}{p} \sum_{t=1}^{p} \left(\frac{1}{n-n_k} \sum_{i \notin M_k} [e_{it} e_{is} - \mathbb{E}(e_{it} e_{is})] \right)^2 \\
&= \quad O_p(n^{-1}).
\end{aligned}
$$

$$
\begin{aligned}
b(k,t) &\leq \quad \frac{1}{p^2} \sum_{t=1}^{p} \|\widetilde{\boldsymbol{l}}_t^{k,d}\|^2 \sum_{t=1}^{p} \left(\mathbb{E}\left(\frac{1}{n-n_k} \sum_{i \notin M_k} e_{it} e_{is} \right) \right)^2 \\
&= \quad \frac{d}{p} \sum_{t=1}^{p} \left(\mathbb{E}\left(\frac{1}{n-n_k} \sum_{i \notin M_k} e_{it} e_{is} \right) \right)^2 \\
&= \quad O_p(p^{-1}).
\end{aligned}
$$

$$
\begin{aligned}
c(k,t) &\leq \quad \frac{1}{p^2} \left\| \sum_{t=1}^{p} \widetilde{\boldsymbol{l}}_t^{k,d} \boldsymbol{l}_t^{0\top} \frac{1}{n-n_k} \sum_{i \notin M_k} \boldsymbol{f}_i^0 e_{is} \right\|^2 \\
&\leq \quad \frac{1}{n-n_k} \left\| \frac{1}{\sqrt{n-n_k}} \sum_{i \notin M_k} \boldsymbol{f}_i^0 e_{is} \right\|^2 \frac{1}{p} \sum_{t=1}^{p} \|\widetilde{\boldsymbol{l}}_t^{k,d}\|^2 \frac{1}{p} \sum_{t=1}^{p} \|\boldsymbol{l}_t^0\|^2 \\
&= \quad O_p(n^{-1}).
\end{aligned}
$$

In summary,

$$
\|\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top} \boldsymbol{l}_s^0\|^2 = O_p(m_{np}^{-1}).
$$

Q.E.D.

LEMMA 2   *Suppose Assumptions 1–4 hold. For any matrix $\boldsymbol{F}$, let $\mathcal{P}_{\boldsymbol{F}} = \boldsymbol{F}(\boldsymbol{F}^\top \boldsymbol{F})^{-1} \boldsymbol{F}^\top$ be its projection matrix. If the random vector $\boldsymbol{v}$ satisfies $\|\boldsymbol{v}\|^2 = O_p(1)$,*

$$
\boldsymbol{v}^\top (\mathcal{P}_{\boldsymbol{L}^0 \boldsymbol{H}^{k,d}} - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}}) \boldsymbol{v} = O_p(m_{np}^{-1/2}).
$$

PROOF: Let $\boldsymbol{D}_{k,0} = \boldsymbol{H}^{k,d\top}\boldsymbol{L}^{0\top}\boldsymbol{L}^0\boldsymbol{H}^{k,d}/p$ and $\boldsymbol{D}_{k,d} = (\widehat{\boldsymbol{L}}^{k,d})^\top\widehat{\boldsymbol{L}}^{k,d}/p$. We have $\|\boldsymbol{D}_{k,d}^{-1}\| = O(1)$ and $\|\boldsymbol{D}_{k,0}^{-1} - \boldsymbol{D}_{k,d}^{-1}\| = O_p(m_{np}^{-1/2})$ (see Lemma 2 of Bai and Ng (2002)). Note that,

$$
\begin{aligned}
& \boldsymbol{v}^\top(\mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}} - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}})\boldsymbol{v} \\
=\ & \frac{1}{p}\boldsymbol{v}^\top\boldsymbol{L}^0\boldsymbol{H}^{k,d}(\boldsymbol{D}_{k,0}^{-1} - \boldsymbol{D}_{k,d}^{-1})\boldsymbol{H}^{k,d\top}\boldsymbol{L}^{0\top}\boldsymbol{v} \\
& -\frac{1}{p}\boldsymbol{v}^\top(\widehat{\boldsymbol{L}}^{k,d} - \boldsymbol{L}^0\boldsymbol{H}^{k,d})\boldsymbol{D}_{k,d}^{-1}(\widehat{\boldsymbol{L}}^{k,d} - \boldsymbol{L}^0\boldsymbol{H}^{k,d})^\top\boldsymbol{v} \\
& -\frac{1}{p}\boldsymbol{v}^\top(\widehat{\boldsymbol{L}}^{k,d} - \boldsymbol{L}^0\boldsymbol{H}^{k,d})\boldsymbol{D}_{k,d}^{-1}\boldsymbol{H}^{k,d\top}\boldsymbol{L}^{0\top}\boldsymbol{v} \\
& -\frac{1}{p}\boldsymbol{v}^\top\boldsymbol{L}^0\boldsymbol{H}^{k,d}\boldsymbol{D}_{k,d}^{-1}(\widehat{\boldsymbol{L}}^{k,d} - \boldsymbol{L}^0\boldsymbol{H}^{k,d})^\top\boldsymbol{v} \\
=\ & I + II + III + IV.
\end{aligned}
$$

We consider the four terms in turn.

$$
\begin{aligned}
I\ &=\ \frac{1}{p}\sum_{s=1}^{p}\sum_{t=1}^{p} v_s \boldsymbol{l}_s^{0\top}\boldsymbol{H}^{k,d}(\boldsymbol{D}_{k,0}^{-1} - \boldsymbol{D}_{k,d}^{-1})\boldsymbol{H}^{k,d\top}\boldsymbol{l}_t^0 v_t \\
.\ &\leq\ \frac{1}{p}\|\boldsymbol{D}_{k,0}^{-1} - \boldsymbol{D}_{k,d}^{-1}\|(\|\sum_{s=1}^{p} v_s \boldsymbol{l}_s^{0\top}\boldsymbol{H}^{k,d}\|^2) \\
&\leq\ \frac{1}{p}\|\boldsymbol{D}_{k,0}^{-1} - \boldsymbol{D}_{k,d}^{-1}\| \cdot \sum_{s=1}^{p}\|\boldsymbol{l}_s^{0\top}\boldsymbol{H}^{k,d}\|^2 \cdot \|\boldsymbol{v}\|^2 \\
&=\ O_p(m_{np}^{-1/2});
\end{aligned}
$$

$$
\begin{aligned}
II\ &=\ \frac{1}{p}\sum_{s=1}^{p}\sum_{t=1}^{p} v_s(\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top}\boldsymbol{l}_s^0)^\top\boldsymbol{D}_{k,d}^{-1}(\widehat{\boldsymbol{l}}_t^{k,d} - \boldsymbol{H}^{k,d\top}\boldsymbol{l}_t^0)v_t \\
&\leq\ \frac{1}{p}\Big\{\sum_{s=1}^{p}\sum_{t=1}^{p}[(\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top}\boldsymbol{l}_s^0)^\top\boldsymbol{D}_{k,d}^{-1}(\widehat{\boldsymbol{l}}_t^{k,d} - \boldsymbol{H}^{k,d\top}\boldsymbol{l}_t^0)]^2\Big\}^{1/2}\|\boldsymbol{v}\|^2 \\
&\leq\ \frac{1}{p}(\sum_{s=1}^{p}\|\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top}\boldsymbol{l}_s^0\|^2)\|\boldsymbol{D}_{k,d}^{-1}\| \cdot O_p(1) \\
&=\ O_p(m_{np}^{-1});
\end{aligned}
$$

$$
\begin{aligned}
III\ &=\ \frac{1}{p}\sum_{s=1}^{p}\sum_{t=1}^{p} x_{is}(\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top}\boldsymbol{l}_s^0)^\top\boldsymbol{D}_{k,d}^{-1}\boldsymbol{H}^{k,d\top}\boldsymbol{l}_t^0 x_{it} \\
&\leq\ \frac{1}{p}\Big\{\sum_{s=1}^{p}\sum_{t=1}^{p}[(\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top}\boldsymbol{l}_s^0)^\top\boldsymbol{D}_{k,d}^{-1}\boldsymbol{H}^{k,d\top}\boldsymbol{l}_t^0]^2\Big\}^{1/2}\|\boldsymbol{v}\|^2 \\
&\leq\ \frac{1}{p}(\sum_{i=1}^{p}\|\widehat{\boldsymbol{l}}_s^{k,d} - \boldsymbol{H}^{k,d\top}\boldsymbol{l}_s^0\|^2)^{1/2}(\sum_{s=1}^{p}\|\boldsymbol{H}^{k,d\top}\boldsymbol{l}_s^0\|^2)^{1/2}\|\boldsymbol{D}_k^{-1}\| \cdot O_p(1) \\
&=\ O_p(m_{np}^{-1/2}).
\end{aligned}
$$

Similarly,

$$
IV = O_p(m_{np}^{-1/2}).
$$

The proof is then complete. *Q.E.D.*

LEMMA 3 *Suppose Assumptions 1–4 hold. For any matrix $\boldsymbol{F}$, denote $W_i(d, \boldsymbol{F}) = \frac{1}{p}\boldsymbol{x}_i^\top(\boldsymbol{I}_d - \mathcal{P}_{\boldsymbol{F}})\boldsymbol{x}_i$. Then, for $1 \le d \le d_0$,*

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k} W_i(d, \widehat{\boldsymbol{L}}^{k,d}) - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k} W_i(d, \boldsymbol{L}^0\boldsymbol{H}^{k,d}) = O_p(m_{np}^{-1/2}).$$

PROOF: Note that, $E\|\boldsymbol{x}_i\|^2 = \sum_{s=1}^{p}\mathbb{E}(x_{is}^2) \le p\max_{i,s}\mathbb{E}(x_{is}^2) = O(p)$. According to Lemma 2,

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}[W_i(d, \widehat{\boldsymbol{L}}^{k,d}) - W_i(d, \boldsymbol{F}^0\boldsymbol{H}^{k,d})]$$
$$= \frac{1}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\boldsymbol{x}_i^\top(\mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}} - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}})\boldsymbol{x}_i$$
$$= O_p(m_{np}^{-1/2}).$$

Q.E.D.

LEMMA 4 *Suppose Assumptions 1–4 hold. For any $d < d_0$, there exists a number $c_d > 0$ such that,*

$$\lim_{n,p\to\infty}\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}[W_i(d, \boldsymbol{L}^0\boldsymbol{H}^{k,d}) - W_i(d_0, \boldsymbol{L}^0)] > c_d.$$

PROOF:

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in M_k}[W_i(d, \boldsymbol{L}^0\boldsymbol{H}^{k,d}) - W_i(d_0, \boldsymbol{L}^0)]$$
$$= \frac{1}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\boldsymbol{x}_i^\top(\mathcal{P}_{\boldsymbol{L}^0} - \mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}})\boldsymbol{x}_i$$
$$= \frac{1}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\boldsymbol{e}_i^\top(\mathcal{P}_{\boldsymbol{L}^0} - \mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}})\boldsymbol{e}_i$$
$$+ \frac{2}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\boldsymbol{e}_i^\top(\mathcal{P}_{\boldsymbol{L}^0} - \mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}})\boldsymbol{L}^0\boldsymbol{f}_i^0$$
$$+ \frac{1}{np}\sum_{k=1}^{K}\sum_{i\in M_k}\boldsymbol{f}_i^{0\top}\boldsymbol{L}^{0\top}(\mathcal{P}_{\boldsymbol{L}^0} - \mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}})\boldsymbol{L}^0\boldsymbol{f}_i^0$$
$$= I + II + III.$$

We have $I \ge 0$ since for each d, $(\mathcal{P}_{\boldsymbol{L}^0} - \mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}})$ is positive semi-definite. For II, note that for each d,

$$\frac{2}{p}\boldsymbol{e}_i^\top(\mathcal{P}_{\boldsymbol{L}^0} - \mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}})\boldsymbol{L}^0\boldsymbol{f}_t^0 = \frac{2}{p}[\boldsymbol{e}_i^\top\boldsymbol{L}^0\boldsymbol{f}_t^0 - \boldsymbol{e}_i^\top\mathcal{P}_{\boldsymbol{L}^0\boldsymbol{H}^{k,d}}\boldsymbol{L}^0\boldsymbol{f}_t^0] = o_p(1),$$

since $\|\frac{1}{\sqrt{p}}\boldsymbol{L}^{0\top}\boldsymbol{e}_i\| = O_p(1)$.

For III, according to Stock and Watson (1998), $\boldsymbol{H}^{k,d}$ converges to a matrix $\boldsymbol{H}_0^d$ with rank d, then

$$
\begin{aligned}
III &= \frac{1}{np} \sum_{k=1}^K \sum_{i \in M_k} \boldsymbol{f}_i^{0\top} \boldsymbol{L}^{0\top} (\mathcal{P}_{\boldsymbol{L}^0} - \mathcal{P}_{\boldsymbol{L}^0 \boldsymbol{H}^{k,d}}) \boldsymbol{L}^0 \boldsymbol{f}_i^0 \\
&\to \frac{1}{np} \sum_{k=1}^K \sum_{i \in M_k} \boldsymbol{f}_i^{0\top} \boldsymbol{L}^{0\top} (\mathcal{P}_{\boldsymbol{L}^0} - \mathcal{P}_{\boldsymbol{L}^0 \boldsymbol{H}_0^d}) \boldsymbol{L}^0 \boldsymbol{f}_i^0 \\
&= \mathrm{Tr}(\frac{1}{p}[\boldsymbol{L}^{0\top}\boldsymbol{L}^0 - \boldsymbol{L}^{0\top}\boldsymbol{L}^0\boldsymbol{H}_0^d(\boldsymbol{H}_0^{d\top}\boldsymbol{L}_0^\top\boldsymbol{L}_0\boldsymbol{H}_0^d)^{-1}\boldsymbol{H}_0^{d\top}\boldsymbol{L}^{0\top}\boldsymbol{L}^0] \cdot \frac{1}{n}\boldsymbol{F}^{0\top}\boldsymbol{F}^0) \\
&> c_d > 0,
\end{aligned}
$$

since $\boldsymbol{H}_0^d$ has rank d and $\lambda_{d_0}(\frac{1}{n}\boldsymbol{F}^{0\top}\boldsymbol{F}^0)$ and $\lambda_{d_0}(\frac{1}{p}\boldsymbol{L}^{0\top}\boldsymbol{L}^0)$ are bounded away from 0. $\hspace{2cm}$ $Q.E.D.$

LEMMA 5 *Suppose Assumptions 1–4 hold. For $1 \le d_1 \le d_0$ and $r < d_2 \le d_{\max}$,*

$$
\boldsymbol{\phi}_{d_1}^\top \widehat{\boldsymbol{\phi}}_{d_2}^k = O_p(m_{np}^{-1}).
$$

PROOF: Let $\widehat{\boldsymbol{\Phi}}_{d_0}^k = (\widehat{\boldsymbol{\phi}}_1^k, \widehat{\boldsymbol{\phi}}_2^k, ..., \widehat{\boldsymbol{\phi}}_{d_0}^k)$ and $\boldsymbol{\Phi}_{d_0} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_{d_0})$. According to Davis-Kahan $\sin\Theta$ theorem (Davis and Kahan, 1970; Yu et al., 2014), there exist a orthogonal matrix $\widehat{\boldsymbol{O}} \in \mathbb{R}^{d_0 \times d_0}$ such that

$$
\|\widehat{\boldsymbol{\Phi}}_{d_0}^k \widehat{\boldsymbol{O}} - \boldsymbol{\Phi}_{d_0}\| \le \frac{\sqrt{d_0}\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{L}^0\boldsymbol{F}_{-M_k}^{0\top}\boldsymbol{F}_{-M_k}^0\boldsymbol{L}^{0\top}\|_{\mathrm{op}}}{\lambda_{d_0} - \lambda_{d_0+1}},
$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm. Note that,

$$
\begin{aligned}
\|\boldsymbol{F}_{-M_k}^{0\top}\boldsymbol{E}_{-M_k}\|_{\mathrm{op}} &= \sup_{\|\boldsymbol{\xi}=1\|} \|\boldsymbol{F}_{-M_k}^{0\top}\boldsymbol{E}_{-M_k}\boldsymbol{\xi}\| \\
&\le \sup_{\|\boldsymbol{\xi}=1\|} \left[(\sum_{s=1}^p \| \sum_{i \notin M_k} \boldsymbol{f}_i^0 e_{is}\|^2)^{1/2} \cdot \|\boldsymbol{\xi}\| \right] \\
&= O_p(\sqrt{np}).
\end{aligned}
$$

Then,

$$
\begin{aligned}
&\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{L}^0\boldsymbol{F}_{-M_k}^{0\top}\boldsymbol{F}_{-M_k}^0\boldsymbol{L}^{0\top}\|_{\mathrm{op}} \\
&= \| \sum_{i \in M_k} \boldsymbol{x}_i\boldsymbol{x}_i^\top + \boldsymbol{L}^0\boldsymbol{F}_{-M_k}^{0\top}\boldsymbol{E}_{-M_k} + \boldsymbol{E}_{-M_k}\boldsymbol{F}_{-M_k}^0\boldsymbol{L}^{0\top} + \boldsymbol{E}_{-M_k}^\top\boldsymbol{E}_{-M_k}\|_{\mathrm{op}} \\
&\le \| \sum_{i \in M_k} \boldsymbol{x}_i\boldsymbol{x}_i^\top \|_{\mathrm{op}} + \|\boldsymbol{L}^0\boldsymbol{F}_{-M_k}^{0\top}\boldsymbol{E}_{-M_k}\|_{\mathrm{op}} + \|\boldsymbol{E}_{-M_k}\boldsymbol{F}_{-M_k}^0\boldsymbol{L}^{0\top}\|_{\mathrm{op}} + \|\boldsymbol{E}_{-M_k}^\top\boldsymbol{E}_{-M_k}\|_{\mathrm{op}} \\
&\le \| \sum_{i \in M_k} \boldsymbol{x}_i\boldsymbol{x}_i^\top \|_{\mathrm{op}} + 2\|\boldsymbol{L}^0\boldsymbol{F}_{-M_k}^{0\top}\|_{\mathrm{op}} \cdot \|\boldsymbol{E}_{-M_k}\|_{\mathrm{op}} + \|\boldsymbol{E}_{-M_k}^\top\boldsymbol{E}_{-M_k}\|_{\mathrm{op}} \\
&= O_p(M_{np}).
\end{aligned}
$$

Consequently,

$$
\|\widehat{\boldsymbol{\Phi}}_{d_0}^k \widehat{\boldsymbol{O}} - \boldsymbol{\Phi}_{d_0}\| = O_p(\frac{M_{np}}{np}) = O_p(m_{np}^{-1}),
$$

Let $\widehat{\boldsymbol{o}}_{d_1}$ be the transpose of the d_1-th row of $\widehat{\boldsymbol{O}}$, we can obtain that,

$$\boldsymbol{\phi}_{d_1}^\top \widehat{\boldsymbol{\phi}}_{d_2}^k = \widehat{\boldsymbol{\phi}}_{d_2}^{k\top}(\boldsymbol{\phi}_{d_1} - \widehat{\boldsymbol{\Phi}}_{d_0}\widehat{\boldsymbol{o}}_{d_1}) + \widehat{\boldsymbol{\phi}}_{d_2}^{k\top}\widehat{\boldsymbol{\Phi}}_{d_0}\widehat{\boldsymbol{o}}_{d_1} = O_p(m_{np}^{-1}).$$

<div align="right">Q.E.D.</div>

LEMMA 6   *Suppose Assumptions 1–4 hold. For $d_0 < d \le d_{\max}$,*

$$\|\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{L}^0\boldsymbol{f}_i^0\|^2 = O_p(pm_{np}^{-2}).$$

PROOF: It is easy to see that $\boldsymbol{L}^0\boldsymbol{f}_i^0 \in span\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_{d_0}\}$. Let $\boldsymbol{L}^0\boldsymbol{f}_i^0 = \sum\limits_{s=1}^{d_0} b_s\boldsymbol{\phi}_s$ with $\sum\limits_{s=1}^{d_0} b_s^2 = \|\boldsymbol{L}^0\boldsymbol{f}_i^0\|^2$. Note that,

$$\begin{aligned}
\|\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{L}^0\boldsymbol{f}_i^0\|^2 &= \|\sum_{s=1}^{d_0} b_s\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{\phi}_s\|^2 \\
&\le \|\boldsymbol{L}^0\boldsymbol{f}_i^0\|^2\sum_{s=1}^{d_0}\|\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{\phi}_s\|^2
\end{aligned}$$

Since $\|\boldsymbol{L}^0\boldsymbol{f}_i^0\|^2 = O_p(p)$, according to Lemma 5, we have that,

$$\|\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{L}^0\boldsymbol{f}_i^0\|^2 = O_p(pm_{np}^{-2}).$$

<div align="right">Q.E.D.</div>

LEMMA 7   *Suppose Assumptions 1–4 hold. Let $\mathrm{Cov}(\boldsymbol{e}_i) = \boldsymbol{\Sigma}_{p,i}$, for $d_0 < d \le d_{\max}$,*

$$\frac{1}{n}\sum_{k=1}^K\sum_{i\in M_k}\sum_{s=1}^p w_s^{k,d}(x_{is} - \widehat{\boldsymbol{f}}_i^{k,d\top}\widehat{\boldsymbol{l}}_s^{k,d})^2 = \frac{1}{n}\sum_{k=1}^K\sum_{i\in M_k}\mathrm{Tr}(\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}}\boldsymbol{\Sigma}_{p,i}) + o_p(1).$$

PROOF: Let $\boldsymbol{H}^{k,d}$ be as defined in Lemma 1 and $\boldsymbol{H}^{k,d-}$ be the generalized inverse matrix of $\boldsymbol{H}^{k,d}$. Moreover, let $\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} = \mathrm{diag}(w_1^{k,d}, w_2^{k,d}, ..., w_p^{k,d})$. It is easy to see that $\sup_s w_s^{k,d} = o_p(1)$. Write $\boldsymbol{x}_i = \boldsymbol{L}^0\boldsymbol{f}_i^0 + \boldsymbol{e}_i = \widehat{\boldsymbol{L}}^{k,d}\boldsymbol{H}^{k,d-}\boldsymbol{f}_i^0 + (\boldsymbol{L}^0\boldsymbol{H}^{k,d} - \widehat{\boldsymbol{L}}^{k,d})\boldsymbol{H}^{k,d-}\boldsymbol{f}_i^0 + \boldsymbol{e}_i$. Since $\|(\boldsymbol{L}^0\boldsymbol{H}^{k,d} - \widehat{\boldsymbol{L}}^{k,d})\boldsymbol{H}^{k,d-}\boldsymbol{f}_i^0\|^2 = O_p(1)$,

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^K\sum_{i\in M_k}\sum_{s=1}^p w_s^{k,d}(x_{is} - \widehat{\boldsymbol{f}}_i^{k,d\top}\widehat{\boldsymbol{l}}_s^{k,d})^2 \\
=\ &\frac{1}{n}\sum_{k=1}^K\sum_{i\in M_k}(\boldsymbol{x}_i - \widehat{\boldsymbol{L}}^{k,d\top}\widehat{\boldsymbol{f}}_i^{k,d})^\top\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}}(\boldsymbol{x}_i - \widehat{\boldsymbol{L}}^{k,d\top}\widehat{\boldsymbol{f}}_i^{k,d}) \\
=\ &\frac{1}{n}\sum_{k=1}^K\sum_{i\in M_k}\boldsymbol{e}_i^\top\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}}\boldsymbol{e}_i + \frac{2}{n}\sum_{k=1}^K\sum_{i\in M_k}\boldsymbol{e}_i^\top\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}}(\boldsymbol{L}^0\boldsymbol{H}^{k,d} - \widehat{\boldsymbol{L}}^{k,d})\boldsymbol{H}^{k,d-}\boldsymbol{f}_i^0 \\
&+ \frac{1}{n}\sum_{k=1}^K\sum_{i\in M_k}((\boldsymbol{L}^0\boldsymbol{H}^{k,d} - \widehat{\boldsymbol{L}}^{k,d})\boldsymbol{H}^{k,d-}\boldsymbol{f}_i^0)^\top\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}}(\boldsymbol{L}^0\boldsymbol{H}^{k,d} - \widehat{\boldsymbol{L}}^{k,d})\boldsymbol{H}^{k,d-}\boldsymbol{f}_i^0
\end{aligned}$$

$$= \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \boldsymbol{e}_i^\top \mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} \boldsymbol{e}_i + o_p(1)$$

$$= \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \text{Tr}(\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} \boldsymbol{\Sigma}_{p,i}) + o_p(1).$$

Q.E.D.

LEMMA 8   *Suppose Assumptions 1–4 hold. Moreover, either Assumption 5.a or 5.b holds. For* $1 \le d_1 \le d_0$ *and* $d_0 + 1 \le d_2 \le d_{\max}$,

$$\frac{1}{\lambda_{d_1} - \lambda_{d_2}} = O_{a.s.}(\frac{1}{np}).$$

Here, λ *can be replaced by* $\widehat{\lambda}$, λ^k, *and* $\widehat{\lambda}^k$.

PROOF: We only prove the result for λ_d and $\widehat{\lambda}_d$ here. The result for λ_d^k and $\widehat{\lambda}_d^k$ can similarly be verified. For λ_s, as $\lambda_d = 0$ for $d > d_0$ and $\lambda_d \ge \lambda_{d_0}$ for $d \le d_0$, the result follows if we can show that,

$$\lambda_{d_0} > Cnp \quad a.s.,$$

with C a positive constant. For any $\boldsymbol{\xi} \in \mathbb{R}^p$ with $\|\boldsymbol{\xi}\|^2 = 1$, we have

$$\boldsymbol{\xi}^\top \boldsymbol{L}^0 \boldsymbol{F}^{0\top} \boldsymbol{F}^0 \boldsymbol{L}^{0\top} \boldsymbol{\xi} \ge \lambda_{d_0}(\boldsymbol{F}^{0\top} \boldsymbol{F}^0) \cdot \|\boldsymbol{L}^{0\top} \boldsymbol{\xi}\|^2 = \lambda_{d_0}(\boldsymbol{F}^{0\top} \boldsymbol{F}^0) \cdot \boldsymbol{\xi}^\top \boldsymbol{L}^0 \boldsymbol{L}^{0\top} \boldsymbol{\xi}.$$

Then, according to Assumption 1, Assumption 2 and Min-Max theorem,

$$\begin{aligned} \lambda_{d_0} &= \min_{dim(\boldsymbol{U})=d_0} \left\{ \max_{\boldsymbol{\xi} \in \boldsymbol{U}, \|\boldsymbol{\xi}\|^2=1} \boldsymbol{\xi}^\top \boldsymbol{L}^0 \boldsymbol{F}^{0\top} \boldsymbol{F}^0 \boldsymbol{L}^{0\top} \boldsymbol{\xi} \right\} \\ &\ge \lambda_{d_0}(\boldsymbol{F}^{0\top} \boldsymbol{F}^0) \cdot \min_{dim(\boldsymbol{U})=d_0} \left\{ \max_{\boldsymbol{\xi} \in \boldsymbol{U}, \|\boldsymbol{\xi}\|^2=1} \boldsymbol{\xi}^\top \boldsymbol{L}^0 \boldsymbol{L}^{0\top} \boldsymbol{\xi} \right\} \\ &= \lambda_{d_0}(\boldsymbol{F}^{0\top} \boldsymbol{F}^0) \cdot \lambda_{d_0}(\boldsymbol{L}^0 \boldsymbol{L}^{0\top}) \\ &\ge Cnp \quad a.s.. \end{aligned}$$

Here $\lambda_{d_0}(\boldsymbol{A})$ represent the d_0-th largest eigenvalue of $\boldsymbol{A}$.

For $\widehat{\lambda}_d$, note that $\boldsymbol{X} = \boldsymbol{F}^0 \boldsymbol{L}^{0\top} + \boldsymbol{E}$, by the Weyl's inequality for singular values (Theorem 3.3.16 of Horn and Johnson (1990)),

$$|\sqrt{\widehat{\lambda}_d} - \sqrt{\lambda_d}| \le \lambda_1^{\frac{1}{2}}(\boldsymbol{E}^\top \boldsymbol{E}).$$

Under Assumption 5.a, $\lambda_1(\frac{1}{n} \boldsymbol{E}^\top \boldsymbol{E}) < 2\sigma^2$. While according to Bao et al. (2015), under Assumption 5.b, $\lambda_1(\frac{1}{n} \boldsymbol{E}^\top \boldsymbol{E})$ is bounded. It follows that,

$$\widehat{\lambda}_{d_0} > (\sqrt{Cnp} - \lambda_1^{\frac{1}{2}}(\boldsymbol{E}^\top \boldsymbol{E}))^2 > C'np \quad a.s.$$

and

$$\widehat{\lambda}_{d_0+1} = O_{a.s.}(\lambda_1(\boldsymbol{E}^\top \boldsymbol{E})) = o_{a.s.}(np).$$

Then,

$$\frac{1}{\widehat{\lambda}_{d_1} - \widehat{\lambda}_{d_2}} \leq \frac{1}{\widehat{\lambda}_{d_0} - \widehat{\lambda}_{d_0+1}} = O_{a.s.}(\frac{1}{np}).$$

                                                                                                        Q.E.D.

LEMMA 9   *Suppose Assumptions 1–4 hold. Moreover, either Assumption 5.a or 5.b holds. For $d_0 < d \leq d_{\max}$,*

$$\sum_{l=d_0+1}^{d} \sum_{k=1}^{K} (\widehat{\lambda}_l - \widehat{\lambda}_l^k) \leq \sum_{l=d_0+1}^{d} \widehat{\lambda}_l$$

and

$$\sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 \leq \sum_{l=d_0+1}^{d} \widehat{\lambda}_l, \quad a.s.$$

PROOF: Let $\widehat{\boldsymbol{\phi}}_l^k = \sum_{s=1}^{p} c_{ls,k} \widehat{\boldsymbol{\phi}}_s$. Note that,

$$\boldsymbol{X}^\top \boldsymbol{X} \widehat{\boldsymbol{\phi}}_l^k = \boldsymbol{X}^\top \boldsymbol{X} \sum_{s=1}^{p} c_{ls,k} \widehat{\boldsymbol{\phi}}_s = \sum_{s=1}^{p} c_{ls,k} \widehat{\lambda}_s \widehat{\boldsymbol{\phi}}_s,$$

and

$$\begin{aligned}
\boldsymbol{X}^\top \boldsymbol{X} \widehat{\boldsymbol{\phi}}_l^k &= \boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} \widehat{\boldsymbol{\phi}}_l^k + \sum_{i \in M_k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k \\
&= \widehat{\lambda}_l^k \widehat{\boldsymbol{\phi}}_l^k + \sum_{i \in M_k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k \\
&= \sum_{s=1}^{p} c_{ls,k} \widehat{\lambda}_l^k \widehat{\boldsymbol{\phi}}_s + \sum_{i \in M_k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k.
\end{aligned}$$

We have

$$\sum_{s=1}^{p} c_{ls,k} (\widehat{\lambda}_s - \widehat{\lambda}_l^k) \widehat{\boldsymbol{\phi}}_s = \sum_{i \in M_k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k.$$

It follows that,

$$\sum_{l=d_0+1}^{d} \sum_{s=1}^{p} c_{ls,k}^2 (\widehat{\lambda}_s - \widehat{\lambda}_l^k)^2 = \sum_{l=d_0+1}^{d} \| \sum_{i \in M_k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k \|^2.$$

Since, for $1 \leq s \leq d_0$ and $d_0 < l \leq d_{\max}$,

$$\frac{1}{\widehat{\lambda}_s - \widehat{\lambda}_l^k} = O_{a.s.}(\frac{1}{np}) \text{ and } \|\boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 \leq \|\boldsymbol{x}_i\|^2 \cdot \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2,$$

we have,

$$\sum_{l=d_0+1}^{d} \sum_{s=1}^{d_0} c_{ls,k}^2 \widehat{\lambda}_s = O_{a.s.}\Big(\sum_{l=d_0+1}^{d} \sum_{i \in M_k} \frac{\|\boldsymbol{x}_i\|^2}{np} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 \Big) = o_{a.s.}\Big(\sum_{l=d_0+1}^{d} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 \Big),$$

Let $\widetilde{\boldsymbol{L}}^{d_0} = \sqrt{p}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_{d_0})$,

$$\sum_{l=d_0+1}^{d} \widehat{\lambda}_l \geq \sum_{l=d_0+1}^{d} \widehat{\boldsymbol{\phi}}_l^{k\top} (I_p - \mathcal{P}_{\widetilde{\boldsymbol{L}}^{d_0}}) \boldsymbol{X}^\top \boldsymbol{X} (I_p - \mathcal{P}_{\widetilde{\boldsymbol{L}}^{d_0}}) \widehat{\boldsymbol{\phi}}_l^k$$

$$
\begin{aligned}
&= \sum_{l=d_0+1}^{d} \widehat{\phi}_l^{k\top} \boldsymbol{X}^\top \boldsymbol{X} \widehat{\phi}_l^k - \sum_{l=d_0+1}^{d} \widehat{\phi}_l^{k\top} \mathcal{P}_{\widetilde{\boldsymbol{L}}^{d_0}} \boldsymbol{X}^\top \boldsymbol{X} \mathcal{P}_{\widetilde{\boldsymbol{L}}^{d_0}} \widehat{\phi}_l^k \\
&= \sum_{l=d_0+1}^{d} \widehat{\lambda}_l^k + \sum_{l=d_0+1}^{d} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\phi}_l^k\|^2 - \sum_{l=d_0+1}^{d} \sum_{s=1}^{d_0} c_{ls,k}^2 \widehat{\lambda}_s \\
&= \sum_{l=d_0+1}^{d} \widehat{\lambda}_l^k + (1 - o_{a.s.}(1)) \sum_{l=d_0+1}^{d} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\phi}_l^k\|^2
\end{aligned}
$$

On the other hand, since $(K-1)\boldsymbol{X}^\top \boldsymbol{X} = \sum_{k=1}^{K} \boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k}$, we have that,

$$
\begin{aligned}
(K-1) \sum_{l=d_0+1}^{d} \widehat{\lambda}_l &= (K-1) \sum_{l=d_0+1}^{d} \widehat{\phi}_l^\top \boldsymbol{X}^\top \boldsymbol{X} \widehat{\phi}_l \\
&= \sum_{k=1}^{K} \sum_{l=d_0+1}^{d} \widehat{\phi}_l^\top \boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} \widehat{\phi}_l \\
&= \sum_{k=1}^{K} \sum_{l=d_0+1}^{d} \widehat{\phi}_l^\top (\boldsymbol{I}_p - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d_0}}) \boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} (\boldsymbol{I}_p - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d_0}}) \widehat{\phi}_l \\
&\quad + \sum_{k=1}^{K} \sum_{l=d_0+1}^{d} \widehat{\phi}_l^\top \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d_0}} \boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d_0}} \widehat{\phi}_l \\
&= \sum_{k=1}^{K} \sum_{l=d_0+1}^{d} \widehat{\lambda}_l^k \cdot \|(\boldsymbol{I}_p - \mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d_0}}) \widehat{\phi}_l\|^2 + \sum_{k=1}^{K} \sum_{l=d_0+1}^{d} \sum_{t=1}^{d_0} \widehat{\lambda}_t^k \|\widehat{\phi}_l^\top \widehat{\phi}_t^k\|^2 \\
&\leq (1 + o(1)) \sum_{k=1}^{K} \sum_{l=d_0+1}^{d} \widehat{\lambda}_l^k.
\end{aligned}
$$

Consequently,

$$
\sum_{l=d_0+1}^{d} \sum_{k=1}^{K} (\widehat{\lambda}_l - \widehat{\lambda}_l^k) \leq (K - (K-1)) \sum_{l=d_0+1}^{d} \widehat{\lambda}_l = \sum_{l=d_0+1}^{d} \widehat{\lambda}_l
$$

and

$$
\sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\phi}_l^k\|^2 \leq \sum_{l=d_0+1}^{d} \widehat{\lambda}_l, \quad a.s.
$$

<div align="right">Q.E.D.</div>

LEMMA 10   *Suppose Assumptions 1–4 and 5.b hold. Then, for $d_0 < d \leq d_{\max}$,*

$$
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}(1 - |\widehat{\phi}_d^\top \widehat{\phi}_d^k|)^2 = o(1).
$$

PROOF: Here we only prove the result for $d = d_0 + 1$. For $d_0 + 1 < d \leq d_{\max}$, the result can be similarly verified. Similar to Lemma 9, set $\widehat{\phi}_d^k = \sum_{s=1}^{p} c_{ds,k} \widehat{\phi}_s$. Then

$$
\sum_{s=1}^{p} c_{ds,k}^2 (\widehat{\lambda}_s - \widehat{\lambda}_d^k)^2 = \| \sum_{i \in M_k} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widehat{\phi}_d^k \|^2.
$$

and

$$\sum_{s=1}^{p} c_{ds,k}^2 (\widehat{\lambda}_s - \widehat{\lambda}_d^k) = \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_d^k\|^2.$$

Write $\xi_1^{k,d} = \sum_{s=1}^{d_0} c_{ds,k}^2$, $\xi_2^{k,d} = \sum_{s=1}^{d_0} c_{ds,k}^2 (\widehat{\lambda}_s - \widehat{\lambda}_d^k)$. It is easy to see that,

$$\frac{1}{K} \sum_{k=1}^{K} \xi_1^{k,d} = o_{a.s.}(1);$$

and, as shown in Lemma 9,

$$\sum_{k=1}^{K} \xi_2^{k,d} = o_{a.s.}(\sum_{k=1}^{K} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_d^k\|^2) = o_{a.s.}(\widehat{\lambda}_d).$$

Note that,

$$
\begin{aligned}
& c_{dd,k}^2 (\widehat{\lambda}_d - \widehat{\lambda}_d^k) + \xi_2^{k,d} \\
= \ & \sum_{s=d+1}^{p} c_{ds,i}^2 (\widehat{\lambda}_d^k - \widehat{\lambda}_s) + \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_d^k\|^2 \\
\geq \ & \sum_{s=d+1}^{p} c_{ds,t}^2 (\widehat{\lambda}_d^k - \widehat{\lambda}_{d+1}) \\
= \ & (1 - \xi_1^{k,d} - c_{dd,k}^2)[(\widehat{\lambda}_d - \widehat{\lambda}_{d+1}) - (\widehat{\lambda}_d - \widehat{\lambda}_d^k)] \\
= \ & (1 - \xi_1^{k,d} - c_{dd,k}^2)(\widehat{\lambda}_d - \widehat{\lambda}_{d+1}) - (1 - \xi_1^{k,d})(\widehat{\lambda}_d - \widehat{\lambda}_d^k) + c_{dd,k}^2 (\widehat{\lambda}_d - \widehat{\lambda}_d^k) \\
\geq \ & (1 - \xi_1^{k,d} - c_{dd,k}^2)(\widehat{\lambda}_d - \widehat{\lambda}_{d+1}) - (\widehat{\lambda}_d - \widehat{\lambda}_d^k) + c_{dd,k}^2 (\widehat{\lambda}_d - \widehat{\lambda}_d^k).
\end{aligned}
$$

We obtain that,

$$\sum_{k=1}^{K} \left[(1 - \xi_1^{k,d} - c_{dd,k}^2)(\widehat{\lambda}_d - \widehat{\lambda}_{d+1}) - (\widehat{\lambda}_d - \widehat{\lambda}_d^k) \right] \leq \sum_{k=1}^{K} \xi_2^{k,d}.$$

Then,

$$\frac{1}{K} \sum_{k=1}^{K} (1 - c_{dd,k}^2) \leq \frac{\frac{1}{K} \sum_{k=1}^{K} [\xi_2^{k,d} + (\widehat{\lambda}_d - \widehat{\lambda}_d^k)]}{\widehat{\lambda}_d - \widehat{\lambda}_{d+1}} + \frac{1}{K} \sum_{k=1}^{K} \xi_1^{k,d}.$$

According to Lemma 9, we have

$$\sum_{k=1}^{K} [\widehat{\lambda}_d - \widehat{\lambda}_d^k] \leq \widehat{\lambda}_d.$$

According to Bao et al. (2015), we have that, under Assumption 5.b, $\widehat{\lambda}_d/n = O_p(1)$ and $\widehat{\lambda}_d - \widehat{\lambda}_{d+1} \sim O_p(n^{1/3})$.

Furthermore, $1/K \leq \sup_k n_k/n = o(n^{-2/3})$. Then,

$$\frac{1}{K} \sum_{k=1}^{K} (1 - c_{dd,k}^2) = O_{a.s.}\left(\frac{\widehat{\lambda}_d}{K(\widehat{\lambda}_d - \widehat{\lambda}_{d+1})} \right) = o_p(1).$$

Note that $0 \leq c_{dd,k}^2 \leq 1$, we have $0 \leq \frac{1}{K} \sum_{k=1}^{K} (1 - c_{dd,k}^2) \leq 1$. Then,

$$\lim_{t \to \infty} \sup_n \mathbb{E}\left(\left| \frac{1}{K} \sum_{k=1}^{K} (1 - c_{dd,k}^2) \right| I_{|\frac{1}{K} \sum_{k=1}^{K} (1 - c_{dd,k}^2)| > t} \right) = 0.$$

According to Theorem 1.8 of Shao (2003),

$$\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K}(1-c_{dd,k}^2)\right) = o(1).$$

Consequently,

$$
\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(1-|\widehat{\phi}_d^{\top}\widehat{\phi}_d^k|)^2 &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(1-|c_{dd,k}|)^2 \\
&\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(1-|c_{dd,k}|) \\
&\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(1-c_{dd,k}^2) = o(1).
\end{aligned}
$$

Q.E.D.

LEMMA 11   *Let $\boldsymbol{e} = (e_1, e_2, ..., e_n)^{\top}$ be a random vector with $\max_i \mathbb{E}(e_i^4) < C$ and $\boldsymbol{a} = (a_1, a_2, ..., a_n)^{\top}$ a random vector that is independent with $\boldsymbol{e}$. Then,*

$$\mathbb{E}(\boldsymbol{a}^{\top}\boldsymbol{e})^4 \leq C\mathbb{E}\|\boldsymbol{a}\|^4.$$

PROOF:

$$
\begin{aligned}
\mathbb{E}(\boldsymbol{a}^{\top}\boldsymbol{e})^4 &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}(a_i a_j a_k a_l e_i e_j e_k e_l)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\mathbb{E}(a_i a_j a_k a_l)\mathbb{E}(e_i e_j e_k e_l) \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\mathbb{E}(a_i a_j a_k a_l)\mathbb{E}(e_i^4)^{1/4}\mathbb{E}(e_j^4)^{1/4}\mathbb{E}(e_k^4)^{1/4}\mathbb{E}(e_l^4)^{1/4} \\
&\leq C\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\mathbb{E}(a_i a_j a_k a_l) \\
&= C\mathbb{E}\left[\left(\sum_{i=1}^{n}a_i\right)^4\right] \\
&\leq C\mathbb{E}\left[\left(\sum_{i=1}^{n}a_i^2\right)^2\right] = C\mathbb{E}\|\boldsymbol{a}\|^4.
\end{aligned}
$$

Q.E.D.

LEMMA 12   *Suppose Assumptions 1–4 and Assumption 5.b hold. For $d_0 < d \leq d_{\max}$, there exist a constant C_0 such that*

$$\widehat{\phi}_d^{k\top}\boldsymbol{\Sigma}_p\widehat{\phi}_d^k \to C_0 \quad a.s..$$

PROOF: For $p \times p$ matrix $\boldsymbol{A}$, let $F^{\boldsymbol{A}}(t) = \frac{1}{p}\sum_{s=1}^{p} \mathbb{1}(\lambda_s(\boldsymbol{A}) \leq t)$ be its empirical spectral distribution (ESD). The Stieltjes transform of a nondecreasing function F is defined by $m_F(z) = \int \frac{1}{t-z}dF(t)$ for all $z \in \mathbb{C}^+$ where $\mathbb{C}^+ = \{z \in \mathbb{C}, Im(z) > 0\}$. Let $\lim_{n \to \infty} p/n = \rho$, by Bai and Zhou (2008), we have, for all k, $F^{\frac{1}{n-n_k}\boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k}} \to F$ $a.s.$ with $m_F(z)$ satisfies $m_F(z) = \int \frac{1}{t(1-\rho-\rho z m_F(z))}dH(t)$.

Let $\boldsymbol{\Sigma}_p = \sum_{t=1}^{p} \tau_t \boldsymbol{v}_t \boldsymbol{v}_t^\top$ and

$$\Phi_p(\lambda, \tau) = \frac{1}{p}\sum_{s=1}^{p}\sum_{t=1}^{p} |\widehat{\boldsymbol{\phi}}_s^{k\top}\boldsymbol{v}_t|^2 \mathbb{1}(\lambda \geq \lambda_s^k)\mathbb{1}(\tau \geq \tau_t).$$

Then, by the procedures in Ledoit and Péché (2011), (7) holds if we can prove that $|\text{Tr}(\boldsymbol{\Sigma}_p^s(\frac{1}{n-n_k}\boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} - z\boldsymbol{I}_p)^{-1}) - \text{Tr}(\boldsymbol{\Sigma}_p^s(\frac{1}{n-n_k}\boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k} - z\boldsymbol{I}_p)^{-1})| \to 0$ for $s = 0, 1, 2, \ldots$.

(7)
$$\Phi_N(\lambda, \tau) \to \int_{-\infty}^{\lambda}\int_{-\infty}^{\tau} \phi(l, t)dH(l)dF(t), \quad a.s.$$

where $\phi(l, t)$ is defined in Ledoit and Péché (2011).

In fact, since $\lambda_1(\boldsymbol{\Sigma}_p^s) = \lambda_1^s(\boldsymbol{\Sigma}_p) < \infty$ and

$$|\text{Tr}(\boldsymbol{\Sigma}_p^s(\frac{1}{n-n_k}\boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} - z\boldsymbol{I}_p)^{-1}) - \text{Tr}(\boldsymbol{\Sigma}_p^s(\frac{1}{n-n_k}\boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k} - z\boldsymbol{I}_p)^{-1})|$$
$$\leq \lambda_1(\boldsymbol{\Sigma}_p^s)|\text{Tr}((\frac{1}{n-n_k}\boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} - z\boldsymbol{I}_p)^{-1}) - \text{Tr}((\frac{1}{n-n_k}\boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k} - z\boldsymbol{I}_p)^{-1})|.$$

we only need to prove that

$$|\text{Tr}((\frac{1}{n-n_k}\boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} - z\boldsymbol{I}_p)^{-1}) - \text{Tr}((\frac{1}{n-n_k}\boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k} - z\boldsymbol{I}_p)^{-1})| \to 0,$$

which follows the facts that

$$F^{\frac{1}{n-n_k}\boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k}} = \text{Tr}((\frac{1}{n-n_k}\boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} - z\boldsymbol{I}_p)^{-1});$$
$$F^{\frac{1}{n-n_k}\boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k}} = \text{Tr}((\frac{1}{n-n_k}\boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k} - z\boldsymbol{I}_p)^{-1});$$
$$F^{\frac{1}{n-n_k}\boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k}} - F^{\frac{1}{n-n_k}\boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k}} \leq \frac{1}{p}\text{rank}(\frac{1}{n-n_k}(\boldsymbol{X}_{-M_k}^\top \boldsymbol{X}_{-M_k} - \boldsymbol{E}_{-M_k}^\top \boldsymbol{E}_{-M_k})) \to 0.$$

According to Weyl's inequality, $\widehat{\lambda}_{d+n_k} \leq \widehat{\lambda}_d^k \leq \widehat{\lambda}_d$. Moreover, by Proposition 1 of Onatski (2010), exists a constant C_λ such that $\lambda_d \to C_\lambda$, $a.s.$ for $d > d_0$ and $d/n \to 0$. This implies that $\widehat{\lambda}_d^k \to C_\lambda$, $a.s.$ for $d_0 < d \leq d_{\max}$. Let (X_p, Y_p) follows Φ_p, (X, Y) follows Φ and $C_0 = E(Y|X = C_\lambda)$. Then,

$$\begin{aligned}
\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{\Sigma}_p\widehat{\boldsymbol{\phi}}_d^k &= \sum_{t=1}^{p} \tau_t |\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{v}_t|^2 \\
&= \frac{\frac{1}{p}\sum_{t=1}^{p} \tau_t |\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{v}_t|^2}{\frac{1}{p}\sum_{t=1}^{p} |\widehat{\boldsymbol{\phi}}_d^{k\top}\boldsymbol{v}_t|^2} \\
&= \frac{\sum_{t=1}^{p} \tau_t P(X_n = \lambda_d^k, Y_n = \tau_t)}{P(X_n = \lambda_d^k)}
\end{aligned}$$

$$= \quad E(Y_n|X_n = \lambda_d^k)$$

$$\rightarrow \quad E(Y|X = C_\lambda) = C_0 \quad a.s..$$

$$Q.E.D.$$

LEMMA 13 *Suppose Assumptions 1–4 and Assumption 5.b hold. For $d_0 < d \le d_{\max}$,*

$$\mathrm{Var}(\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i) \rightarrow 0.$$

PROOF: Similar to Lemma 10, we only prove the result for $d = d_0 + 1$. Denote $\boldsymbol{X}_{-M_{k,l}}$ be the submatrix of $\boldsymbol{X}$ with the rows in M_k and M_l being removed, $\widehat{\lambda}_d^{k,l}$'s, and $\widehat{\boldsymbol{\phi}}_k^{k,l}$'s be the corresponding eigenvalues and eigenvectors.

Note that $\boldsymbol{e}_i$ is independent with $\widehat{\boldsymbol{\phi}}_d^k$ for $i \in M_k$, we have

$$\mathbb{E}[\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i] = \mathbb{E}[\widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{\Sigma}_p \widehat{\boldsymbol{\phi}}_d^k] = C_0 + o(1).$$

Then, let $i \ne j$ and $i \in M_k$, $j \in M_l$,

$$\mathrm{Cov}[\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i, \boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^l \widehat{\boldsymbol{\phi}}_d^{l\top} \boldsymbol{e}_j]$$

$$= \quad \mathbb{E}[\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i \boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^l \widehat{\boldsymbol{\phi}}_d^{l\top} \boldsymbol{e}_j] - \mathbb{E}[\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i] \mathbb{E}[\boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^l \widehat{\boldsymbol{\phi}}_d^{l\top} \boldsymbol{e}_j]$$

$$= \quad \mathbb{E}[\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i \boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^{k,l} \widehat{\boldsymbol{\phi}}_d^{k,l\top} \boldsymbol{e}_j] + \mathbb{E}[\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i (\boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^l \widehat{\boldsymbol{\phi}}_d^{l\top} \boldsymbol{e}_j - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^{k,l} \widehat{\boldsymbol{\phi}}_d^{k,l\top} \boldsymbol{e}_j)] - C_0^2 + o(1)$$

$$= \quad (I) + (II) - C_0^2 + o(1).$$

For (I), we have

$$(I) \quad = \quad \mathbb{E}\{\mathbb{E}[\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i \boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^{k,l} \widehat{\boldsymbol{\phi}}_d^{k,l\top} \boldsymbol{e}_j | (\boldsymbol{F}^0, \boldsymbol{L}^0, \boldsymbol{E}_{-M_k})]\}$$

$$= \quad \mathbb{E}\{\boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^{k,l} \widehat{\boldsymbol{\phi}}_d^{k,l\top} \boldsymbol{e}_j \mathbb{E}[\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i | (\boldsymbol{F}^0, \boldsymbol{L}^0, \boldsymbol{E}_{-M_k})]\}$$

$$= \quad \mathbb{E}[\boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^{k,l} \widehat{\boldsymbol{\phi}}_d^{k,l\top} \boldsymbol{e}_j \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{\Sigma}_p \widehat{\boldsymbol{\phi}}_d^k]$$

$$= \quad C_0 \mathbb{E}[\widehat{\boldsymbol{\phi}}_d^{k,l\top} \boldsymbol{\Sigma}_p \widehat{\boldsymbol{\phi}}_d^{k,l}] + o(1)$$

$$= \quad C_0^2 + o(1).$$

For (II), without loss of generality, we assume that $0 \le \widehat{\boldsymbol{\phi}}_d^{l\top} \widehat{\boldsymbol{\phi}}_d^{k,l} \le 1$. Otherwise we use $-\widehat{\boldsymbol{\phi}}_d^{k,l}$ to replace $\widehat{\boldsymbol{\phi}}_d^{k,l}$. Then, $2 \le \|\widehat{\boldsymbol{\phi}}_d^l + \widehat{\boldsymbol{\phi}}_d^{k,l}\|^2 \le 4$.

Note that $\boldsymbol{e}_j$ is independent with $\widehat{\boldsymbol{\phi}}_d^l$ and $\widehat{\boldsymbol{\phi}}_d^{k,l}$, by Lemma 11,

$$\mathbb{E}\left((\widehat{\boldsymbol{\phi}}_d^l + \widehat{\boldsymbol{\phi}}_d^{k,l})^\top \boldsymbol{e}_j\right)^4 \le C\mathbb{E}\left[\|\widehat{\boldsymbol{\phi}}_d^l + \widehat{\boldsymbol{\phi}}_d^{k,l})\|^4\right] < \infty,$$

and

$$\mathbb{E}\left(\boldsymbol{e}_j^\top (\widehat{\boldsymbol{\phi}}_d^l - \widehat{\boldsymbol{\phi}}_d^{k,l})\right)^4 \le C\mathbb{E}\left[\|\widehat{\boldsymbol{\phi}}_d^l - \widehat{\boldsymbol{\phi}}_d^{k,l}\|\right]^4.$$

It follows that,

$$
\begin{aligned}
(II) \quad \le \quad & \left[\mathbb{E}(\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i)^2\right]^{1/2}\left[\mathbb{E}(\boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^l \widehat{\boldsymbol{\phi}}_d^{l\top} \boldsymbol{e}_j - \boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^{k,l} \widehat{\boldsymbol{\phi}}_d^{k,l\top} \boldsymbol{e}_j)^2\right]^{1/2} \\
\le \quad & \left[\mathbb{E}(\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i)^2\right]^{1/2}\left\{\mathbb{E}\left[\boldsymbol{e}_j^\top (\widehat{\boldsymbol{\phi}}_d^l + \widehat{\boldsymbol{\phi}}_d^{k,l})\boldsymbol{e}_j^\top(\widehat{\boldsymbol{\phi}}_d^l - \widehat{\boldsymbol{\phi}}_d^{k,l})\right]^2\right\}^{1/2} \\
\le \quad & \left[\mathbb{E}(\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i)^2\right]^{1/2}\left[\mathbb{E}(\boldsymbol{e}_j^\top(\widehat{\boldsymbol{\phi}}_d^l + \widehat{\boldsymbol{\phi}}_d^{k,l}))^4\right]^{1/4}\left[\mathbb{E}(\boldsymbol{e}_j^\top(\widehat{\boldsymbol{\phi}}_d^l - \widehat{\boldsymbol{\phi}}_d^{k,l}))^4\right]^{1/4} \\
\le \quad & C\left(\left[\mathbb{E}\|\widehat{\boldsymbol{\phi}}_d^l - \widehat{\boldsymbol{\phi}}_d^{k,l}\|^4\right]^{1/4}\right. \\
= \quad & C'\left[\mathbb{E}(1 - |\widehat{\boldsymbol{\phi}}_d^{l\top} \widehat{\boldsymbol{\phi}}_d^{k,l}|)^2\right]^{1/4}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
& \mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i \in M_k} \boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i\right) \\
= \quad & \frac{1}{n^2}\sum_{k=1}^{K}\sum_{i \in M_k} \mathrm{Var}(\boldsymbol{e}_i^\top \boldsymbol{\phi}_d^k \boldsymbol{\phi}_d^{k\top} \boldsymbol{e}_i) + \frac{1}{n^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i \in M_k}\sum_{\substack{j \ne i \\ j \in M_l}} \mathrm{Cov}(\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_d^k \widehat{\boldsymbol{\phi}}_d^{k\top} \boldsymbol{e}_i, \boldsymbol{e}_j^\top \widehat{\boldsymbol{\phi}}_d^l \widehat{\boldsymbol{\phi}}_d^{l\top} \boldsymbol{e}_j)] \\
\le \quad & \frac{C}{n} + \frac{C'}{n^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i \in M_k}\sum_{\substack{j \ne i \\ j \in M_l}}\left[\mathbb{E}(1 - |\widehat{\boldsymbol{\phi}}_d^{l\top}\widehat{\boldsymbol{\phi}}_d^{k,l}|)^2\right]^{1/4} \\
= \quad & \frac{C}{n} + \frac{C'}{n^2}\sum_{k=1}^{K}\sum_{l \ne k}\sum_{i \in M_k}\sum_{j \in M_l}\left[\mathbb{E}(1 - |\widehat{\boldsymbol{\phi}}_d^{l\top}\widehat{\boldsymbol{\phi}}_d^{k,l}|)^2\right]^{1/4} \\
= \quad & \frac{C}{n} + \frac{C'}{n^2}\sum_{k=1}^{K}\sum_{l \ne k} n_k n_l\left[\mathbb{E}(1 - |\widehat{\boldsymbol{\phi}}_d^{l\top}\widehat{\boldsymbol{\phi}}_d^{k,l}|)^2\right]^{1/4} \\
\le \quad & \frac{C}{n} + \frac{C'K\sup_k n_k}{n^2}\sum_{k=1}^{K} n_k \frac{1}{K}\sum_{l \ne k}\left[\mathbb{E}(1 - |\widehat{\boldsymbol{\phi}}_d^{l\top}\widehat{\boldsymbol{\phi}}_d^{k,l}|)^2\right]^{1/4} \\
= \quad & \frac{C}{n} + o(C'K\sup_k n_k/n) \\
= \quad & o(1).
\end{aligned}
$$

Here, the last last equalities follow from Lemma 10 and Assumption 4. Q.E.D.

LEMMA 14 *Suppose Assumptions 1–4 hold. Moreover, either Assumption 5.a or 5.b holds. Then, for $d_0 < d \le d_{\max}$,*

$$
\frac{1}{n}\sum_{l=d_0+1}^{d}\sum_{k=1}^{K}\sum_{i \in M_k}\|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 < \frac{2}{n}\sum_{k=1}^{K}\sum_{i \in M_k}\mathrm{Tr}((\mathcal{Q}_{\widehat{\boldsymbol{L}}^k,d} - \mathcal{Q}_{\widehat{\boldsymbol{L}}^k,d_0})\boldsymbol{\Sigma}_{p,i}) + o_p(1).
$$

PROOF: On the one hand, when Assumption 1 to 3 and Assumption 5.a hold, we have,

$$
\begin{aligned}
\mathrm{Tr}((\mathcal{Q}_{\widehat{\boldsymbol{L}}^k,d} - \mathcal{Q}_{\widehat{\boldsymbol{L}}^k,d_0})\boldsymbol{\Sigma}_{p,i}) \quad = \quad & \sum_{s=1}^{p}(w_s^{k,d} - w_s^{k,d_0})\sigma^2 \\
= \quad & \sigma^2[\mathrm{Tr}(\mathcal{P}_{\widehat{\boldsymbol{L}}^k,d}) - \mathrm{Tr}(\mathcal{P}_{\widehat{\boldsymbol{L}}^k,d_0})] \\
= \quad & (d - d_0)\sigma^2.
\end{aligned}
$$

While, according to Lemma 9,

$$\frac{1}{n} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 \leq \frac{1}{n} \sum_{l=d_0+1}^{d} \widehat{\lambda}_l, \quad a.s.$$

By Weyl's inequality, we have

$$\frac{1}{n} \sum_{l=d_0+1}^{d} \widehat{\lambda}_l \leq \frac{d-d_0}{n} \lambda_1(\boldsymbol{E}^\top \boldsymbol{E}) < 2(d-d_0)\sigma^2 = \frac{2}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \mathrm{Tr}((\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} - \mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d_0}})\boldsymbol{\Sigma}_{p,i}) \quad a.s..$$

On the other hand, when Assumption 1 to 3 and Assumption 5.b hold, note that,

$$\frac{1}{n} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 = \frac{1}{n} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \left(\|\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 + 2\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_l^k \widehat{\boldsymbol{\phi}}_l^{k\top} \boldsymbol{L}^0 \boldsymbol{f}_i^0 + \|\boldsymbol{f}_i^{0\top} \boldsymbol{L}^{0\top} \widehat{\boldsymbol{\phi}}_l^k\|^2 \right).$$

Since $\mathbb{E}\|\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 = O(1)$ and $\mathbb{E}\|\boldsymbol{f}_i^{0\top} \boldsymbol{L}^{0\top} \widehat{\boldsymbol{\phi}}_l^k\|^2 = O(pm_{np}^{-2})$, we have,

$$|\frac{1}{np} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_l^k \widehat{\boldsymbol{\phi}}_l^{k\top} \boldsymbol{L}^0 \boldsymbol{f}_i^0|$$

$$\leq \frac{1}{np} \sum_{k=d_0+1}^{d} (\sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2)^{1/2} (\sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{f}_i^{0\top} \boldsymbol{L}^{0\top} \widehat{\boldsymbol{\phi}}_l^k\|^2)^{1/2}$$

$$= O_p(p^{1/2} m_{np}^{-1}) = o_p(1);$$

and

$$\frac{1}{n} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{f}_i^{0\top} \boldsymbol{L}^{0\top} \widehat{\boldsymbol{\phi}}_l^k\|^2 = O_p(pm_{np}^{-2}) = o_p(1).$$

In addition, according to Lemma 13, for $d_0 + 1 \leq l \leq d_{\max}$,

$$\mathbb{E}(\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_l^k \widehat{\boldsymbol{\phi}}_l^{k\top} \boldsymbol{e}_i) \to C_0;$$

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \widehat{\boldsymbol{\phi}}_l^{k\top} \boldsymbol{\Sigma}_p \widehat{\boldsymbol{\phi}}_l^k \to C_0 \quad a.s.;$$

and

$$\mathrm{Var}(\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_l^k \widehat{\boldsymbol{\phi}}_l^{k\top} \boldsymbol{e}_i) \to 0.$$

Thus

$$\frac{1}{n} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{e}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 = \sum_{l=d_0+1}^{d} C_0 + o_p(1).$$

Note that $\mathcal{P}_{\widehat{\boldsymbol{L}}^{k,d}} = \sum_{l=1}^{d} \widehat{\boldsymbol{\phi}}_l^k \widehat{\boldsymbol{\phi}}_l^{k\top}$, we have,

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \mathrm{Tr}((\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} - \mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d_0}})\boldsymbol{\Sigma}_{p,i}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \sum_{l=d_0+1}^{d} \widehat{\boldsymbol{\phi}}_l^{k\top} \mathrm{diag}(\boldsymbol{\Sigma}_p) \widehat{\boldsymbol{\phi}}_l^k.$$

Thus,

$$\frac{1}{n} \sum_{l=d_0+1}^{d} \sum_{k=1}^{K} \sum_{i \in M_k} \|\boldsymbol{x}_i^\top \widehat{\boldsymbol{\phi}}_l^k\|^2 - \frac{2}{n} \sum_{k=1}^{K} \sum_{i \in M_k} \mathrm{Tr}((\mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d}} - \mathcal{Q}_{\widehat{\boldsymbol{L}}^{k,d_0}})\boldsymbol{\Sigma}_{p,i})$$

$$= \quad \widehat{\phi}_l^{k\top}(\boldsymbol{\Sigma}_p - 2\mathrm{diag}(\boldsymbol{\Sigma}_p))\widehat{\phi}_l^k + o_p(1)$$

$$< \quad 0 + o_p(1),$$

since $2\mathrm{diag}(\boldsymbol{\Sigma}_p) - \boldsymbol{\Sigma}_p$ is positive definite. $Q.E.D.$