Journal of Multivariate Analysis 133 (2015) 140–159

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

High dimensional mean–variance optimization through factor analysis

Binbin Chen [a], Shih-Feng Huang [b], Guangming Pan [a,*]

[a] Nanyang Technological University, 637371, Singapore

[b] National University of Kaohsiung, Taiwan

ARTICLE INFO

Article history:
Received 22 November 2012
Available online 28 September 2014

AMS subject classification:
62H12

Keywords:
Factor model
Optimal portfolio allocation
Mean–variance optimization

ABSTRACT

A factor analysis-based approach for estimating high dimensional covariance matrix is proposed and is applied to solve the mean–variance portfolio optimization problem in finance. The consistency of the proposed estimator is established by imposing a factor model structure with a relative weak assumption on the relationship between the dimension and the sample size. Numerical results indicate that the proposed estimator outperforms the plug-in, linear shrinkage and bootstrap-corrected approaches.

1. Introduction

The mean–variance (MV) portfolio optimization procedure introduced by Markowitz [20,21] is the cornerstone of modern portfolio theory for optimal portfolio construction, asset allocation and investment diversification. In this procedure, portfolio optimizers attempt to maximize portfolio expected return for a given amount of portfolio risk, or equivalently minimize risk for a given level of expected return, by carefully choosing the proportions of various assets [20–22,24,16]. The closed-form formula of the MV optimization problem consists of the number of assets, denoted by p, the expected returns of the assets, denoted by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$, and the corresponding covariance matrix, denoted by Σ_y. Since $\boldsymbol{\mu}$ and Σ_y are actually unknown, a natural idea is to replace them by the sample mean vector, $\bar{\mathbf{y}}$, and sample covariance matrix, S_y, of the training sample. However, this plug-in estimator is found to have substantial bias and no longer optimal, especially when p is large. Frankfurter, Phillips and Seagle [12] and Jobson and Korkie [14] reported that the plug-in portfolio can perform worse than an equally weighted portfolio. Hence, doubts remained on the MV optimization procedure and the procedure had not been widely accepted by the investment community in the 1980s. Michaud [25] termed this phenomenon the “Markowitz optimization enigma” and called the MV optimizers “estimation-error maximizers”.

The bias of the plug-in estimator is shown to be caused by an extremely large number of p [23] and the estimation of Σ_y [17]. Many studies are devoted to correcting this bias by using different approaches, see [2] and references therein. Among others, by random matrix techniques Bai, Liu and Wong [2] proposed a bootstrap-corrected estimator to correct the bias of estimating Σ_y^{-1} by S_y^{-1} and obtained the optimal allocation and expected return by a bootstrap procedure. The bootstrap-corrected estimator is shown to be capable of adequately solving the MV optimization problem when $p < n$, where n denotes the number of observations. However, it is hard to extend their approach to the case when $p > n$ since the bootstrap-corrected estimator still relies on S_y, the estimator of Σ_y. Therefore, this study proposes a factor analysis-based estimator of a high dimensional covariance matrix when p is proportional to n or $p > n$, and applies the proposed estimator to solve the MV optimization problem.

* Corresponding author.
E-mail address: gmpan@ntu.edu.sg (G. Pan).

http://dx.doi.org/10.1016/j.jmva.2014.09.006

To tackle the situation of $p > n$, various methods were proposed to estimate the population covariance matrix Σ_y. One is the linear shrinkage method, which shrinks the sample covariance matrix to a multiple of the identity. By minimizing the expectation of the difference of the shrinkage estimator against the true covariance matrix under some criteria (e.g. the Frobenius norm), the shrinkage intensity can be obtained [18]. The linear shrinkage method was further improved through the Rao–Blackwell theorem by Chen, Wiesel and Hero [6] for the normally distributed data. Another approach is to reduce the dimensionality by imposing some structure on the data, such as sparsity and compound symmetric. Consistent estimators of Σ_y have been derived by imposing the sparsity on the population covariance matrix in a large amount of literature (see [4]). Imposing the sparsity directly on the covariance matrix, however, is inappropriate for the MV portfolio optimization problem, as in practice it is rare that the correlations between asset returns are zero.

Instead, imposing factor model structure on data is one of the most frequently used and effective ways to achieve dimension-reduction. In the Arbitrage Price Theory of Ross [30], a fundamental assumption is that asset returns follow a factor structure. Besides, factor models are also used for business cycle analysis, forecasting diffusion indices and consumer theory, see [1] and papers therein. By assuming that a few factors can completely capture the cross-sectional risks, the number of parameters in covariance matrix estimation can be significantly reduced. With the purpose of estimating the population covariance, imposing the factor model structure on the data set was also investigated by Fan, Fan and Lv [9] for strict factor models and Fan, Liao and Mincheva [10] for approximate factor models. By letting the common factors be observed in advance, the convergence rates to Σ_y^{-1} were investigated under the Frobenius norm and the spectral norm in these two papers, respectively. For high dimensional data, we may face the problem that the common factors are unknown, and therefore we have to estimate the common factors first. In this paper, we study the MV portfolio optimization problem by assuming the data have the factor model structure with unobserved factors.

The rest of this paper is organized as follows. In Section 2, the MV portfolio optimization problem, the factor model and the proposed estimator are introduced. The asymptotic properties of the proposed estimator are established. In Section 3, a comparison study of the proposed method, plug-in, bootstrap-corrected estimator and shrinkage estimator is conducted via simulations. One example of real data is analyzed by using the proposed factor analysis-based approach in Section 4. To facilitate the presentation, proofs, tables and figures are given in the Appendix.

2. Models and estimation method

In this section, we first briefly introduce the MV portfolio optimization problem. Let $\mathbf{y} = (y_1, \ldots, y_p)^\top$ denote the returns of p financial assets with mean $\boldsymbol{\mu}$ and covariance matrix Σ_y. An investor is usually interested in forming a portfolio $\mathbf{w}^\top\mathbf{y}$, where $\mathbf{w} = (w_1, \ldots, w_p)^\top$ is the portfolio allocation, and determining the best allocation of these assets that can maximize the expected return $R = \mathbf{w}^\top\boldsymbol{\mu}$ under certain risk level for high dimensional data. For example,

$$R = \max_{\mathbf{w}} \mathbf{w}^\top\boldsymbol{\mu} \quad \text{subject to } \mathbf{w}^\top\mathbf{1}_p \le 1 \text{ and } \mathbf{w}^\top\Sigma_y\mathbf{w} \le \sigma_0^2, \tag{2.1}$$

where $\mathbf{1}_k$ represents the k-dimensional vector of ones, σ_0^2 is a given risk level, the initial capital is assumed to be less than or equal to 1, and $\mathbf{w}^\top\Sigma_y\mathbf{w}$ is the quadratic risk measure of the portfolio. The well-known analytical solution of (2.1) is

1. if $\sigma_0\mathcal{B} \le \sqrt{\mathcal{A}}$, then $R = \sigma_0\sqrt{\mathcal{A}}$ and $\mathbf{w} = \frac{\sigma_0}{\sqrt{\mathcal{A}}}\Sigma_y^{-1}\boldsymbol{\mu}$;
2. if $\sigma_0\mathcal{B} > \sqrt{\mathcal{A}}$, then $R = \frac{\mathcal{B}}{\mathcal{C}} + b(\mathcal{A} - \frac{\mathcal{B}^2}{\mathcal{C}})$ and $\mathbf{w} = \frac{1}{\mathcal{C}}\Sigma_y^{-1}\mathbf{1}_p + b\Big(\Sigma_y^{-1}\boldsymbol{\mu} - \frac{\mathcal{B}}{\mathcal{C}}\Sigma_y^{-1}\mathbf{1}_p\Big)$;

where $b = \sqrt{\frac{\mathcal{C}\sigma_0^2-1}{\mathcal{A}\mathcal{C}-\mathcal{B}^2}}$, $\mathcal{A} = \boldsymbol{\mu}^\top\Sigma_y^{-1}\boldsymbol{\mu}$, $\mathcal{B} = \mathbf{1}_p^\top\Sigma_y^{-1}\boldsymbol{\mu}$ and $\mathcal{C} = \mathbf{1}_p^\top\Sigma_y^{-1}\mathbf{1}_p$, see Proposition 2.1 in [2]. Note that different solutions were derived under the settings different from (2.1), see for example, [15,8,19,7,31,26,32]. This study focuses on (2.1) to discuss the MV optimization problem.

Suppose the historical p-dimensional returns are $\mathbf{y}_1, \ldots, \mathbf{y}_n$ over an investment period. With the purpose of estimating the optimal allocation $\mathbf{w}$ and the optimal return R, the conventional estimators of $\boldsymbol{\mu}$ and Σ_y are the sample mean, denoted by $\bar{\mathbf{y}}$, and sample covariance matrix S_y, respectively. If one plugs $\bar{\mathbf{y}}$ and S_y directly into the theoretical representation of $\mathbf{w}$ and R and denotes the estimator of R by $R_p = \hat{\mathbf{w}}^\top\bar{\mathbf{y}}$, then the plug-in return R_p, however, will overestimate the theoretic optimal return, especially when the dimension p increases (see Fig. 1). The poor performance and counter-intuitive asset allocation of the plug-in estimator are referred to as the "Markowitz optimization enigma".

The objective of this study is to correct the bias of the plug-in estimator in the situation of p proportional to n or $p > n$. In the following, the considered factor model and the proposed factor analysis-based estimator are introduced in Sections 2.1 and 2.2, respectively. The corresponding asymptotic properties of the proposed estimator are established in Section 2.3.

2.1. The factor model

We assume the data have the following factor model structure:

$$y_{it} = \mu_i + \mathbf{b}_i^\top\mathbf{f}_t + u_{it}, \tag{2.2}$$

where y_{it} is the observed data for the ith cross section at time t ($i = 1, \ldots, p$, $t = 1, \ldots, n$), μ_i is the mean of y_{it}, $\mathbf{b}_i$ is a $K \times 1$ vector of factor loadings, $\mathbf{f}_t$ is a $K \times 1$ vector of unobserved common factors, $K \le p$ is the number of factors and u_{it} is

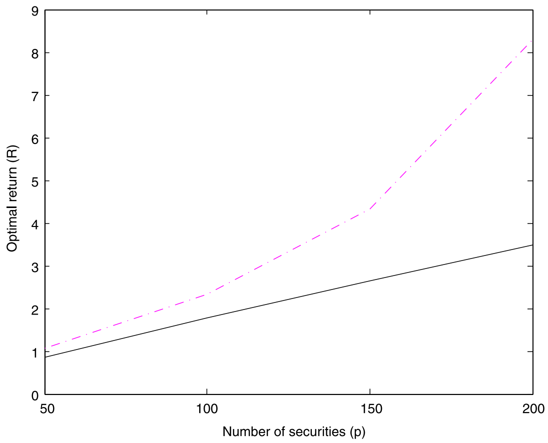

Fig. 1. The plug-in returns R_p (dash line) and the theoretical optimal returns R (solid line) for different numbers of assets.

the idiosyncratic error component of y_{it}. Let

$$\mathbf{y}_t = (y_{1t}, \dots, y_{pt})^\top, \qquad \mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_p), \qquad \mathbf{u}_t = (u_{1t}, \dots, u_{pt})^\top.$$

Then the model (2.2) can be written as a p-dimensional time series with n observations:

$$\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{B}^\top \mathbf{f}_t + \mathbf{u}_t. \tag{2.3}$$

In practical applications, p can be thought of as the number of assets or stocks which can be of the same order of n or even much larger than n. In contrast, the number of factors, K, can be much smaller. There is a large body of literature on how to determine K (see [3,13,27]).

In this paper, we assume that the number of common factors K is known and fixed. The factor loadings $\mathbf{B}$ is deterministic. The unobserved factors $\{\mathbf{f}_t,\ t \geq 1\}$ are *i.i.d.* with mean $\mathbf{0}_K$ and covariance matrix Σ_f, where $\mathbf{0}_k$ is the k-dimensional vector with all elements being zero. The error vector $\mathbf{u}_t,\ t = 1, \dots, n$ are *i.i.d.* with mean $\mathbf{0}_p$ and covariance matrix Σ_u, and are independent with $\{\mathbf{f}_t,\ t \geq 1\}$.

We further assume that the error covariance matrix Σ_u belongs to the following matrix class:

$$\begin{aligned} \mathbb{U}_p\Big(M, \kappa(p,n)\Big) = \Bigg\{ & \Sigma : \Sigma = (\sigma_{ij}) \text{ is a } p \times p \text{ symmetric positive definite matrix}, \\ & \left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right) \cdot \max_{i \leq p} \sum_{j=1}^{p} I\left(|\sigma_{ij}| > M\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right)\right) \leq \kappa(p,n), \\ & \max_{i \leq p} \sum_{j=1}^{p} |\sigma_{ij}| I\left(|\sigma_{ij}| \leq M\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right)\right) \leq \kappa(p,n), \Bigg\}, \end{aligned}$$

where M is a constant and $\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}} \leq \kappa(p,n) \to 0$. This class of matrices can be better understood with the aid of the following examples.

Example 1 (*Sparsity*)**.** Let $\Sigma_u = (\sigma_{ij})$. Define

$$\gamma_n = \max_{i \leq p} \sum_{j=1}^{p} I(\sigma_{ij} \neq 0).$$

Assume that $\gamma_n\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right) = o(1)$. Then $\Sigma_u \in \mathbb{U}_p\Big(C, \kappa(p,n)\Big)$ with $\kappa(p,n) = C\gamma_n\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right)$.

This type of the sparse error covariance Σ_u was employed in [10] when estimating the covariance matrix of $\mathbf{y}_t$ in (2.3) by assuming that $\{\mathbf{f}_t, t \geq 1\}$ are known.

Example 2 (*AR(1) Error Vector*)**.** Suppose that the error vectors $\mathbf{u}_t,\ t = 1, \dots, n$ are i.i.d. copies of $\mathbf{u} = (u_1, \dots, u_p)^\top$. Consider the AR(1) model:

$$u_i = \psi u_{i-1} + \sqrt{1 - \psi^2} e_i, \quad i = 2, \dots, p,$$

where $u_1 = e_1, 0 < |\psi| < 1, e_i \sim N(0,1),\ i = 1, \ldots, n$ are i.i.d. and e_i is independent of u_j for $j < i$. Then the covariance matrix of $\mathbf{u}$ is $\Sigma_u = (|\psi|^{|i-j|})$. An easy calculation implies $\Sigma_u \in \mathbb{U}_p\Big(C, \kappa(p,n)\Big)$ with $\kappa(p,n) = C_\psi(\log p + \log n)\Big(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\Big)$, where C_ψ is a constant depending on ψ.

Indeed, any population covariance matrix possessing the form of $\Sigma_u = (|\psi|^{|i-j|})$ with $0 < |\psi| < 1$ belongs to the class $\mathbb{U}_p\Big(C, \kappa(p,n)\Big)$. It can be verified as follows. For simplicity, consider the case of $i = 1$ and the remaining cases can be evaluated similarly. Note that $\sigma_{1j} = |\psi|^{j-1}$. The condition to ensure

$$\sigma_{1j} = |\psi|^{j-1} > M\Big(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\Big) \tag{2.4}$$

is

$$j - 1 < \frac{1}{\log|\psi|}\left(\log M\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right)\right)$$

which further implies that

$$j < \frac{1}{2|\log|\psi||}\Big(\log p + \log n\Big). \tag{2.5}$$

In other words, (2.5) implies (2.4). Now take $\kappa(p,n) = C_\psi(\log p + \log n)(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}})$ with $C_\psi = \frac{c}{|\log|\psi||}$ and an appropriate constant $c \ge 1/2$. Therefore the first condition in the definition of the matrix class $\mathbb{U}_p\Big(M, \kappa(p,n)\Big)$ is satisfied. Moreover, in view of (2.4) and (2.5)

$$\sum_{j=1}^{p} |\sigma_{1j}| I\left(|\sigma_{1j}| \le M\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right)\right) \le \sum_{j=c_p}^{p} |\psi|^{j-1} \le \frac{|\psi|^{c_p}}{1-|\psi|},$$

where $c_p = \Big[\frac{c}{|\log|\psi||}\Big(\log p + \log n\Big)\Big]$ with $[a]$ denoting the integer not larger than a. Note that

$$|\psi|^{c_p} = e^{-|\log|\psi||\Big[\frac{c}{|\log|\psi||}\Big(\log p + \log n\Big)\Big]}.$$

On the other hand

$$\kappa(p,n) > C_\psi \frac{\log p}{\sqrt{p}} \ge C_\psi e^{-\Big(\frac{1}{2}\log p - \log\log p\Big)}.$$

It follows that

$$\kappa(p,n) \ge \sum_{j=1}^{p} |\sigma_{1j}| I\left(|\sigma_{1j}| \le M\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right)\right).$$

Therefore the second condition in the definition of the matrix class $\mathbb{U}_p\Big(M, \kappa(p,n)\Big)$ is satisfied. These imply that $\Sigma_u = (|\psi|^{|i-j|})$ belongs to the matrix class $\mathbb{U}_p\Big(M, \kappa(p,n)\Big)$.

2.2. The factor analysis-based estimator

The estimation of $\mathbf{B}$ is not unique since model (2.3) is unchanged if we transfer $(\mathbf{B}^\top, \mathbf{f}_t)$ to $(\mathbf{B}^\top \Upsilon, \Upsilon^{-1}\mathbf{f}_t)$ for any $K \times K$ invertible matrix Υ. However the linear space spanned by the columns of $\mathbf{B}^\top$ is uniquely determined by (2.3). Hence we can use the normalization that $\frac{1}{p}\mathbf{B}\mathbf{B}^\top = \mathbf{I}_K$.

Estimating factor loadings $\mathbf{B}$ and the common factors $\mathbf{f}_t$ through principal components is one of the effective method for strong factor models [3,1]. The vector $\boldsymbol{\mu}$ is estimated by sample mean, $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{y}} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_t$ and $\widehat{\boldsymbol{\mu}} \triangleq (\widehat{\mu}_1, \ldots, \widehat{\mu}_p)$. The method of principal components minimizes

$$\min_{\mathbf{B},\mathbf{F}} \frac{1}{np} \sum_{i=1}^{p}\sum_{t=1}^{n} (y_{it} - \widehat{\mu}_i - \mathbf{b}_i^\top \mathbf{f}_t)^2 = \min_{\mathbf{B},\mathbf{F}} \frac{1}{np} \mathrm{tr}\Big(\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}_n^\top - \mathbf{B}^\top\mathbf{F}\Big)\Big(\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}_n^\top - \mathbf{B}^\top\mathbf{F}\Big)^\top$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)$, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ and $\mathrm{tr}\,\mathbf{A}$ stands for the trace of the matrix $\mathbf{A}$. The columns of the estimated factor loadings $\widehat{\mathbf{B}}^\top$ are $\sqrt{p}$ times eigenvectors corresponding to the K largest eigenvalues of the $p \times p$ matrix $\frac{1}{np}(\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}_n^\top)(\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}_n^\top)^\top$. Hence $\frac{1}{p}\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top = \mathbf{I}_K$. We further estimate the common factors and the error terms as:

$$\widehat{\mathbf{F}} = \frac{1}{p}\widehat{\mathbf{B}}(\mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}_n^\top) = (\widehat{\mathbf{f}}_1, \ldots, \widehat{\mathbf{f}}_n) \quad \text{and} \quad \widehat{\mathbf{U}} = \mathbf{Y} - \bar{\mathbf{y}}\mathbf{1}_n^\top - \widehat{\mathbf{B}}^\top\widehat{\mathbf{F}} = (\widehat{u}_{ij}). \tag{2.6}$$

The initial error covariance matrix estimator is constructed as

$$\widehat{\Sigma}_u = \frac{1}{n}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top = \Big(\frac{1}{n}\sum_{t=1}^{n}\widehat{u}_{it}\widehat{u}_{jt}\Big) \triangleq (\widehat{\sigma}_{ij}).$$

We next use a thresholding technique to improve the initial estimation of the error covariance matrix. There are two types of thresholding techniques: universal thresholding and adaptive thresholding. As pointed out by Cai and Liu [4], using a universal constant as the thresholding value may not capture the variability of the individual estimation. Hence we below use the adaptive thresholding estimator [4,10], which is given by

$$\widehat{\Sigma}_u^\tau = (\widehat{\sigma}_{ij}^\tau), \quad \widehat{\sigma}_{ij}^\tau = \widehat{\sigma}_{ij} I(|\widehat{\sigma}_{ij}| \geq \omega_n \sqrt{\widehat{\theta}_{ij}}),$$

$$\widehat{\theta}_{ij} = \frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}\widehat{u}_{jt} - \widehat{\sigma}_{ij})^2,$$

where $\omega_n = \varpi\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right)$ and ϖ is a tuning parameter. Here ω_n is based on the convergence rate of $\widehat{\sigma}_{ij}$ to σ_{ij} from Lemma 2.1 and $\widehat{\theta}_{ij}$ is the estimate of the variance of $\widehat{\sigma}_{ij}$. The value of tuning parameter ϖ can be fixed or be determined by cross-validation.

From model (2.3), we note that

$$\Sigma_y = \mathbf{B}^\top \Sigma_f \mathbf{B} + \Sigma_u.$$

By the Sherman–Morrison–Woodbury formula,

$$\Sigma_y^{-1} = \Sigma_u^{-1} - \Sigma_u^{-1}\mathbf{B}^\top\Big(\Sigma_f^{-1} + \mathbf{B}\Sigma_u^{-1}\mathbf{B}^\top\Big)^{-1}\mathbf{B}\Sigma_u^{-1}. \tag{2.7}$$

We estimate the covariance of $\mathbf{f}_t$ as $\widehat{\Sigma}_f = \frac{1}{n}\widehat{\mathbf{F}}\widehat{\mathbf{F}}^\top$. Therefore, by employing the estimator $\widehat{\mathbf{B}}$, $\widehat{\Sigma}_f$ and thresholding estimator $\widehat{\Sigma}_u^\tau$, we obtain the resulting estimator

$$\widehat{\Sigma}_y = \widehat{\mathbf{B}}^\top \widehat{\Sigma}_f \widehat{\mathbf{B}} + \widehat{\Sigma}_u^\tau,$$

and

$$\widehat{\Sigma}_y^{-1} = (\widehat{\Sigma}_u^\tau)^{-1} - (\widehat{\Sigma}_u^\tau)^{-1}\widehat{\mathbf{B}}^\top\Big(\widehat{\Sigma}_f^{-1} + \widehat{\mathbf{B}}(\widehat{\Sigma}_u^\tau)^{-1}\widehat{\mathbf{B}}^\top\Big)^{-1}\widehat{\mathbf{B}}(\widehat{\Sigma}_u^\tau)^{-1}. \tag{2.8}$$

Note that $\widehat{\mathbf{B}}^\top$ may be not the estimator of the true $\mathbf{B}^\top$ but an estimator of $\mathbf{B}^\top\Upsilon$ for some $K \times K$ invertible matrix Υ. Hence $\widehat{\Sigma}_f$ may be not the estimator of the true covariance Σ_f either but a estimator of $\Upsilon^{-1}\Sigma_f(\Upsilon^{-1})^\top$. But this phenomenon will not affect the estimation of Σ_y since Σ_y is invariant under $(\mathbf{B}^\top, \mathbf{F}) \mapsto (\mathbf{B}^\top\Upsilon, \Upsilon^{-1}\mathbf{F})$, e.g. $\Sigma_y = \mathbf{B}^\top\Upsilon(\Upsilon^{-1}\Sigma_f(\Upsilon^{-1})^\top)\Upsilon^\top\mathbf{B} + \Sigma_u$.

2.3. Asymptotic theory

The following assumptions are imposed for establishing the asymptotic theory:

Assumption A. The number of factors K is known and fixed. The dimension p is a function of n, e.g. $p = p(n)$, satisfying the condition $\sqrt{\frac{\log p}{\log n}} \leq \alpha$ for some positive constant α as $n \to \infty$.

Assumption B. $\{\mathbf{u}_t,\ t = 1, 2, \ldots\}$ are i.i.d. p-dimensional errors vector with mean $\mathbf{0}_p$ and covariance matrix $\Sigma_u = (\sigma_{ij})$. The covariance matrix Σ_u belongs to the class $\mathbb{U}_p\Big(M, \kappa(p,n)\Big)$ with M being a constant and $\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}} \leq \kappa(p,n) \to 0$ as $n \to \infty$. There exist constants $c_0, c_1, \delta > 0$ such that $E|u_{it}|^{4(2\alpha+2)+\delta} < c_1$, $E|u_{it}u_{jt}|^{2(4\alpha+2)+\delta} \leq c_1$ for all $i, j \leq p$, $c_0 < \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) < c_1$ and $c_0 < \mathrm{var}(u_{it}u_{jt}) = \sigma_{ij} < c_1$ for all $i, j \leq p$.

Assumption C. The factor loading $\mathbf{B} = (b_{ki})$ is a deterministic $K \times p$ matrix with $\max_{k,i}|b_{ki}| \leq c_2$ for some $c_2 > 0$ and $\lambda_{\min}(\frac{1}{p}\mathbf{B}\mathbf{B}^\top) \geq c_0$.

Assumption D. $\{\mathbf{f}_t,\ t = 1, 2, \ldots\}$ are i.i.d. K-dimensional unobserved factors with mean 0_K and covariance matrix $\Sigma_f = (\rho_{ij})$ and $\lambda_{\min}(\Sigma_f) \geq c_0$. The factors $\{f_t,\ t = 1, \ldots, n\}$ are independent of $\{u_t,\ t = 1, \ldots, n\}$.

Determining the number of factors K in factor model has been extensively studied in the literature. The condition $\sqrt{\frac{\log p}{\log n}} \leq \alpha$ reveals that the dimension p can be in the polynomial order of the sample size n, e.g. $p \sim n^\alpha$. For example, if $p/n \to$ some positive constant then $\alpha = 1$.

Here, we put some moment assumptions on the error u_{it} rather than the exponential-type tail distribution, imposed in [10].

The following lemma illustrates the difference between the estimated error terms and the true error terms. This allows us to estimate the covariance matrix by using thresholding when direct observations of error terms are not available.

Lemma 2.1. *Under Assumptions A–D, we have*

(i)

$$P\Big(\max_{i\le p}\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}-u_{it*})^2\le C\Big(\frac{1}{p}+\frac{\log p}{n}\Big)\Big)=1-o(1),$$

(ii)

$$P\Big(\max_{i,j}\Big|\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}\widehat{u}_{jt}-u_{it*}u_{jt*})\Big|\le C\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)\Big)=1-o(1),$$

where (and in what follows) C is a positive constant independent of p, n but may change in different places, $o(1)$ is a term converging to zero as $n\to\infty$ and $u_{it*}=u_{it}-\bar{u}_i$ with $\bar{u}_i=\frac{1}{n}\sum_{t=1}^{n}u_{it}$.

If $\left(\frac{1}{p}+\frac{\log p}{n}\right)\to 0$ as $n\to\infty$, Lemma 2.1 establishes an overall result of the difference between the estimated errors and true errors, e.g. $\max_{i\le p}\frac{1}{n}\mathbf{e}_i^{\mathsf T}(\widehat{\mathbf{U}}-\mathbf{U}_*)(\widehat{\mathbf{U}}-\mathbf{U}_*)^{\mathsf T}\mathbf{e}_i\xrightarrow{\text{i.p.}}0$, where $\mathbf{U}_*=(u_{it*})$ and $\xrightarrow{\text{i.p.}}$ represents "converges in probability". To establish the asymptotic properties of the thresholding estimator $\widehat{\Sigma}_u^{\tau}$, Fan, Liao and Mincheva [10] also need the condition that $\max_{i\le p,t\le n}|\widehat{u}_{it}-u_{it*}|\le C$ in probability for some positive constant C (and this condition is satisfied when the factors are observed as they proved). However, it is hard to check such a condition in the framework of unobserved factors. Here, we manage to derive the asymptotic properties of $\widehat{\Sigma}_u^{\tau}$ without such a condition if we let $n/p^2\to 0$, as $n\to\infty$. In what follows, $\|\mathbf{A}\|$ denotes the spectral norm which is defined as $\|\mathbf{A}\|=\lambda_{\max}^{1/2}(\mathbf{A}\mathbf{A}^{\mathsf T})$.

Theorem 2.1. *Assume that $n/p^2\to 0$. Under Assumptions A–D, we then have*

$$P\Big(\|\widehat{\Sigma}_u^{\tau}-\Sigma_u\|\le C\kappa(p,n)\Big)=1-o(1),$$

for some positive constant C.

Our interest lies in high dimensional data, especially when the dimension is close to or larger than the number of observations. Hence, the assumption $n/p^2\to 0$ is easily satisfied. We observe from here that $\widehat{\Sigma}_u$ is a consistent estimator of Σ_u under the spectral norm with the convergence rate $\kappa(p,n)$. Thanks to this consistency, the consistent estimator of the inverse of the population covariance matrix Σ_y is illustrated in the following theorem, relying on which the estimators of the optimized portfolio allocation and optimal return are obtained.

Theorem 2.2. *Assume that $n/p^2\to 0$. Under Assumptions A–D, we then have*

$$P\Big(\|(\widehat{\Sigma}_y^{\tau})^{-1}-\Sigma_y^{-1}\|\le C\kappa(p,n)\Big)=1-o(1),$$

for some positive constant C independent of p, n.

Remark 2.1. After submission we have been drawn attention to [11] which has some overlap with our work. Below we list some key differences between our work and [11].

First, part (c) of Assumption 2 in [11] requires exponentially decaying tails on the model residuals. This strong condition implies that all moments exist. In applications to daily returns this is likely to be violated. In contrast, we only need finite moment assumption.

Second, when considering the population covariance matrix $\Sigma_u=|\psi|^{|i-j|}$, the convergence rate in Theorems 2.1 and 2.2 is better than that in Theorem 1 of [11]. Specifically speaking, the convergence rate in Theorem 2.1 is

$$\|\hat{\Sigma}_u^{\tau}-\Sigma_u\|=O_p(\kappa(p,n)) \tag{2.9}$$

where $\kappa(p,n)=C_\psi(\log p+\log n)(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}})$ with $C_\psi=\frac{c}{|\log|\psi||}$ (see Example 2). However Theorem 1 of [11] yields

$$\|\hat{\Sigma}_u^{\tau}-\Sigma_u\|=O_p(\omega_T^{(1-q)}m_p), \tag{2.10}$$

where $\omega_T=\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}$ and $m_p=\max_{i\le p}\sum_{j\le p}|\sigma_{ij}|^q$ with $0\le q<1$. When $q=0$ (2.10) yields $\|\hat{\Sigma}_u^{\tau}-\Sigma_u\|=O_p(p(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}))$. It is easy to verify that $\frac{p(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}})}{\kappa(p,n)}\to\infty$. Moreover when $q>0$,

$$\omega_T^{(1-q)}m_p=O\left(\left(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\right)^{1-q}\right).$$

Again one may verify that

$$\frac{\left(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\right)^{1-q}}{\kappa(p,n)}\to\infty.$$

Summarizing the above we conclude that

$$\frac{\omega_T^{(1-q)}m_p}{\kappa(p,n)}\to\infty,$$

which implies that the convergence rate in (2.9) is better than that in (2.10).

Finally, the covariance matrix of the factor loading in [11] is assumed to be an identity matrix while we assume that it is a positive definite symmetric matrix.

For the MV portfolio optimization procedure, we plug the estimators $\bar{\mathbf{y}}$ and $\widehat{\Sigma}_y^{-1}$ into the theoretical representation $\mathcal{A}, \mathcal{B}, \mathcal{C}$ in the solution of (2.1). The asymptotic properties are shown in the following theorem.

Theorem 2.3. *Assume $n/p^2\to 0$. Then under Assumptions A–D, we have*

(i)

$$P\Big(\Big|\frac{\bar{\mathbf{y}}^{\top}\widehat{\Sigma}_y^{-1}\bar{\mathbf{y}}}{p}-\frac{\boldsymbol{\mu}^{\top}\Sigma_y^{-1}\boldsymbol{\mu}}{p}\Big|\le C\kappa(p,n)\Big)=1-o(1),$$

(ii)

$$P\Big(\Big|\frac{\bar{\mathbf{y}}^{\top}\widehat{\Sigma}_y^{-1}\mathbf{1}_p}{p}-\frac{\boldsymbol{\mu}^{\top}\Sigma_y^{-1}\mathbf{1}_p}{p}\Big|\le C\kappa(p,n)\Big)=1-o(1),$$

(iii)

$$P\Big(\Big|\frac{\mathbf{1}_p^{\top}\widehat{\Sigma}_y^{-1}\mathbf{1}_p}{p}-\frac{\mathbf{1}_p^{\top}\Sigma_y^{-1}\mathbf{1}_p}{p}\Big|\le C\kappa(p,n)\Big)=1-o(1),$$

where C is a positive constant independent of p, n.

Remark 2.2. Although we let K be fixed, our result can be directly extended to the case when K increases slowly along with the dimension p. If K tends to infinity, then $\lambda_{\max}(\frac{1}{p}\mathbf{B}\mathbf{B}^T)$ may be not bounded any more. Instead it is bounded by K depending on p. As a consequence, $\kappa(p,n)$ in Lemma 2.1 and Theorem 2.1 should be replaced by $K\kappa(p,n)$ and $\kappa(p,n)$ in Theorems 2.2 and 2.3 replaced by $K^4\kappa(p,n)$ by a careful inspection on the proofs. Also, the condition $K^4\kappa(p,n)\to 0$ is required. In addition, the constant M and $\kappa(p,n)$ involved in the definition of the class $\mathbb{U}_p\Big(M,\kappa(p,n)\Big)$ should be replaced by KM and $K\kappa(p,n)$ respectively.

3. Simulation study

3.1. Factor model structure data

3.1.1. Simulations

In this section, several simulation scenarios are constructed to compare the performance of the factor analysis-based estimator (FACT) with the plug-in approach (PLN), the bootstrap-corrected estimator (BSP) proposed by Bai, Liu and Wong [2] when $p<n$, and the modified linear shrinkage estimator (SRK) proposed by Chen, Wiesel and Hero [6] for both $p<n$ and $p>n$. The p-dimensional data are generated from the model:

$$\begin{aligned}&\mathbf{y}_t=\boldsymbol{\mu}+\mathbf{B}'\mathbf{f}_t+\mathbf{u}_t,\quad t=1,\dots,n,\\&\mathbf{f}_t=\Sigma_f^{1/2}\mathbf{s}_t,\qquad \mathbf{u}_t=\Sigma_u^{1/2}\mathbf{z}_t,\quad t=1,\dots,n,\end{aligned}\tag{3.1}$$

where $\mathbf{s}_t$ and $\mathbf{z}_t$ are respectively K-dimensional and p-dimensional random vectors with *i.i.d.* components. We generate the components of $\mathbf{s}_t$ and $\mathbf{z}_t$ from two kinds of distributions: the standard normal distribution and the standardized Gamma(4, 2). As the mean $\boldsymbol{\mu}$ and factor loadings $\mathbf{B}$ are deterministic, we first randomly generate the elements of $\boldsymbol{\mu}$ from $U(0.8, 1.2)$ and the elements of $\mathbf{B}$ from $N(0, 1)$, respectively, and then fix $\boldsymbol{\mu}$ and $\mathbf{B}$ for each replication. Here, $U(a, b)$ denotes the uniform distribution on the interval $[a, b]$.

In the simulations, set $\Sigma_f=(\rho_{ij})$ with

$$\rho_{ii}=\phi_i\quad\text{and}\quad\rho_{ij}=0.5^{|i-j|}\sqrt{\rho_{ii}\cdot\rho_{jj}},\quad i,j=1,\dots,K,$$

where $\phi_i,\ i = 1, \ldots, p$ are independently generated from $U(0.8, 1.3)$. Set

$$\boldsymbol{\varrho} = (\varrho_1, \ldots, \varrho_{10}) = (0.156, 0.153, 0.073, 0.149, 0.096, 0.113, 0.125, 0.055, 0.016, 0.064),$$

$p_k,\ k = 1, \ldots, 9$ are the integers most close to $p \times \varrho_k$ and $p_{10} = p - \sum_{k=1}^{9} p_k$. The covariance of the error $\Sigma_u = (\sigma_{ij})$ is a 10 block diagonal matrix defined as

$$\Sigma_u = \mathbf{D}_u^{1/2} \operatorname{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_{10}) \mathbf{D}_u^{1/2},$$

where $\mathbf{D}_u = \operatorname{diag}(\varphi_1, \ldots, \varphi_p)$ and $\mathbf{V}_k$ is a $p_k \times p_k$ Toeplitz matrix with the first row being $(1, 0.5, 0.25, 0.125,\ \mathbf{0}_{p_k-4}^{\mathsf{T}})$ when $p_k > 4$ and being $(1, \ldots, 0.5^{p_k-1})$ when $p_k \le 4$. In the simulations, we generate $\varphi_i,\ i = 1, \ldots, p$, independently from $U(0.8, 1.3)$. Here we would point out that Σ_u is a kind of special Toeplitz matrix and most of entries of $\mathbf{V}_k$ are equal to zero except four of them on each row. Therefore, one can verify that it satisfies Assumption B with C being any positive constant independent of n and $\kappa(p; n) = 4\left(\frac{1}{\sqrt{p}} + \sqrt{\frac{\log p}{n}}\right)$.

Our results focus on the large dimensional factor model structure data. To inspect the impact caused by the sample size and the dimension, the following three scenarios are considered to compare the performance of FACT with SRK, PIN and BSP.

(A) If p is much smaller than n, we set $(p, n) = (200, 405), (400, 605), (600, 805)$.
(B) If p is close to but smaller than n, we set $(p, n) = (400, 405), (600, 605), (800, 805)$.
(C) If p is bigger than n, we set $(p, n) = (600, 405), (800, 405), (800, 605)$.

The true number of factors, denoted by K_0, is set to be $K_0 = \lfloor p/100 \rfloor$, which increases slowly along with the dimension p, where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x. In the proposed FACT method, the number of factors K is determined by Onatski's [27] method. We carry out 30 replications for each (p, n)-combination and for each setting of $\mathbf{s}_t$ and $\mathbf{z}_t$.

For each fixed (p, n), we summarize the simulation steps as follows:

(1) Generate $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$, $\mathbf{B} = (b_{ki})$, ϕ_i and $\varphi_i,\ i = 1, \ldots, p$, respectively as $\mu_i \sim U(0.8, 1.2)$, $b_{ki} \sim N(0, 1)$, ϕ_i and $\varphi_i \sim U(0.8, 1.3)$.
(2) Generate the ρ_{ij}, $\mathbf{D}_f$ and $\mathbf{V}_k,\ k = 1, \ldots, 10$, as illustrated above to obtain Σ_f and Σ_u.
(3) Calculate the true optimal return of (2.1) and denote it by R.
(4) Generate the elements of $\mathbf{s}_t$ and $\mathbf{z}_t,\ t = 1, \ldots, n$, independently from $N(0, 1)$.
(5) Calculate $\mathbf{y}_t = \boldsymbol{\mu} + \mathbf{B}^{\mathsf{T}}\mathbf{f}_t + \mathbf{u}_t,\ t = 1, \ldots, n$.
(6) Apply Onatski's [27] method to determine the number of factors K. Then, adopt the factor analysis approach to obtain the estimators of $\mathbf{B}$, $\mathbf{F}$ and $\mathbf{U}$. Set $\omega_n = 0.75\left(\frac{1}{p} + \frac{\log p}{n}\right)^{1/2}$ to derive the thresholding estimators $\widehat{\Sigma}_u^{\tau}$ and $\widehat{\Sigma}_y^{-1}$.
(7) Plug-in the sample mean $\bar{\mathbf{y}}$ and $\widehat{\Sigma}_y^{-1}$ to derive the factor analysis-based estimated optimal return R_f. Appealing to the Shrinkage-based estimator, plug-in approach (when $p < n$) and bootstrap-corrected estimator (when $p < n$), we calculate the estimated optimal return R_s, R_p and R_b, respectively.
(8) Calculate the mean square errors (MSEs) $M_i = (R_f - R)^2,\ i = f, s, p, b$.
(9) Repeat the steps (4)–(8) for 30 replications to obtain 30 values of M_f, M_s, M_p and M_b.

To conduct the simulations for the elements of $\mathbf{s}_t$ and $\mathbf{z}_t$ being generated from the standardized Gamma(4, 2), we repeat the steps (1)–(9) with step (4) revised as "Generate the elements of $\mathbf{s}_t$ and $\mathbf{z}_t,\ t = 1, \ldots, n$, independently from the standardized Gamma(4, 2)". We graph the four mean square errors (or M_f and M_s for $p > n$) for each (p, n)-combination and each distribution of $\mathbf{s}_t$ and $\mathbf{z}_t$.

3.1.2. Results

In Figs. 2–4, the blue, red, green and black curves represent the MSEs $M_i,\ i = f, s, p, b$ of FACT, SRK, PLN and BSP, respectively, for each simulation time. In each figure, the upper panel corresponds to normal distribution that $\mathbf{s}_t$ and $\mathbf{z}_t$ are generated from while the lower panel corresponds to the standardized Gamma(4, 2).

Fig. 2 presents the performance of the four approaches when p is much smaller than n. As observed, the MSEs of FACT and BSP are smaller than SRK and PLN. Meanwhile, the accuracies for FACT and BSP are roughly the same when n is large enough. This reveals that FACT outperforms SRK and PLN significantly, but shares the same good performance as BSP. Fig. 3 compares the performances when p is close to but smaller than n. As we can observe, both FACT and SRK significantly outperform PLN and BSP. The average MSEs for PLN and BSP are much larger than those for FACT and SRK. This is expected since both PLN and BSP rely heavily on the property of the sample covariance matrix $\mathbf{S}_y$ which is very unstable when p is close to n. On the other hand, FACT also retains the remarkable accuracy and outperforms SRK especially when p increases. Fig. 4 corresponds to the case $p > n$. In this scenario, PLN and BSP are not applicable, so we only show the MSEs of FACT and SRK. It is observed that FACT still outperforms SRK remarkably. From Figs. 2–4, although the true number of factors K_0 increases slowly along with the dimension p, the proposed FACT method has significant improvements on solving the MV portfolio optimization problem, especially when p is close to or larger than n.

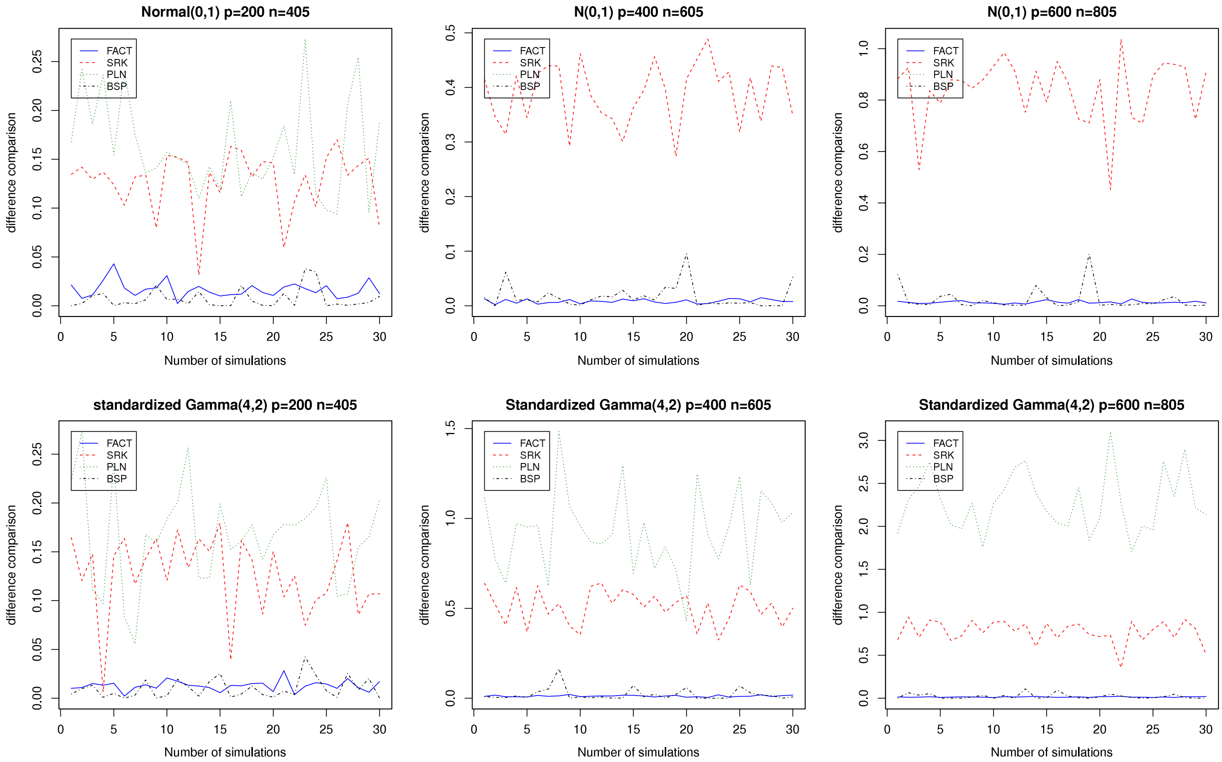

Fig. 2. This figure shows the performance when the data are generated from the factor model (3.1) and the dimension p is much smaller than the sample size n. The mean square errors (MSEs) of factor analysis-based approach (blue solid line), linear shrinkage method (red dotted line), plug-in method (green dashed line) and bootstrap-corrected approach (black dashed line) are plotted for each simulation time. The elements of $\mathbf{s}_t$ and $\mathbf{z}_t$, $t = 1, \ldots, n$ are generated from the normal distribution (upper panel) and the standardized Gamma(4, 2) (lower panel). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Compound symmetric correlation

This section is intended to compare the performances of our methods with the others when the data are not generated from the exact factor model. We generate the p-dimensional data from the following multivariate model:

$$\mathbf{y}_t = \Sigma_y^{1/2}\mathbf{z}_t + \boldsymbol{\mu}, \quad t = 1, 2, \ldots, n, \tag{3.2}$$

where $\mathbf{z}_t$ is a p-dimensional random vector with *i.i.d.* components. We also generate the components of $\mathbf{z}_t$ from two kinds of distributions: the standard normal distribution and the standardized Gamma(4, 2). The elements of $\boldsymbol{\mu}$ are generated from $U(0.8, 1.2)$, then kept fixed. The covariance matrix $\Sigma_y = \mathbf{DVD}$, where $\mathbf{D} = \mathrm{diag}(\sqrt{\mu_1}, \ldots, \sqrt{\mu_p})$ and $\mathbf{V}$ is a compound symmetric matrix with the diagonal elements being equal to 1 and the off diagonal elements being equal to ρ. Noting Σ_y can be decomposed as

$$\Sigma_y = \rho\mathbf{D}\mathbf{1}_p\mathbf{1}_p^{\mathrm{T}}\mathbf{D} + (1-\rho)\mathbf{D}^2,$$

we can view the multivariate model (3.2) as a "factor-type" model with the number of factors $K = 1$. In the simulation, we set $\rho = 0.5$.

We set three (p, n)-combinations, $(p, n) = (200, 405)$, $(400, 405)$ and $(800, 405)$, to inspect the impact caused by the relationship of the sample size and the dimension. 30 replications are conducted for each (p, n)-combination and each setting of $\mathbf{z}_t$. The MSEs, M_f, M_s, M_p and M_b, are plotted in Fig. 5 to compare the performance of the four approaches.

In Fig. 5, the upper panel corresponds to normal distribution that $\mathbf{z}_t$ are generated from while the lower panel corresponds to the standardized Gamma(4, 2). The same story happens when the data are generated from the multivariate model (3.2) instead of the factor model (3.1). When p is much smaller than n, FACT and BSP roughly share the same good performance and both outperform SRK and PLN. When p is close to n or $p > n$, the advantage of FACT will show gradually. However, comparing with the first example when the data are generated from the factor model, factor analysis-based approach will lead to a bit bigger MSEs when the data are generated from the multivariate model.

In conclusion, by imposing a factor model structure, the factor analysis-based approach outperforms the linear shrinkage method, plug-in method and bootstrap-corrected approach with a very weak assumption on the relationship of the dimension and the sample size.

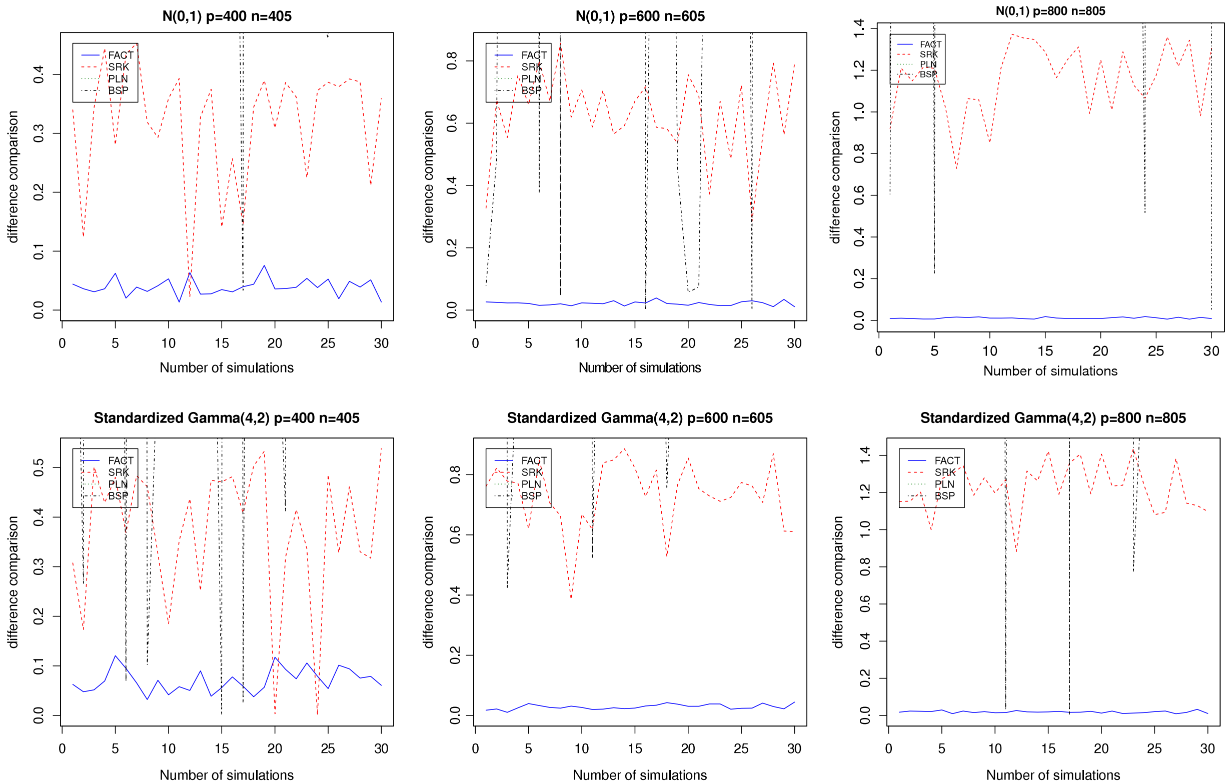

Fig. 3. This figure shows the performance when the data are generated from the factor model (3.1) and the dimension p is close to but smaller than the sample size n. The mean square errors (MSEs) of the factor analysis-based approach (blue solid line), linear shrinkage method (red dotted line), plug-in method (green dashed line) and bootstrap approach (black dashed line) are plotted for each simulation time. However, the MSEs for the plug-in method and bootstrap-corrected approach are so large that they jump out of the bound of the plot. The elements of $\mathbf{s}_t$ and $\mathbf{z}_t$, $t = 1, \ldots, n$ are generated from the normal distribution (upper panel) and the standardized Gamma(4, 2) (lower panel). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Case study

In this section, data from S&P 500 are studied to compare the performances of the factor analysis-based approach with the linear shrinkage, plug-in and bootstrap-corrected approaches. Daily closing bid prices from 1 Jan 2009 to 31 Dec 2011 are collected for 505 stocks in S&P 500 and these 505 stocks are grouped into sectors: consumer discretionary, consumer staples, energy, financials, health care, industrials, information technology, materials, telecommunication services and utilities, respectively. The numbers of stocks in these sectors are 86, 41, 40, 82, 52, 62, 69, 30, 9 and 34 respectively. The data can be downloaded from Wharton Research Data Services.

To detect the impact of the number of stocks involved in the portfolio, we select p ($p = 100, 203, 302, 405$) stocks from the S&P 500 database randomly but with roughly equal weight in each sector. We are also interested in whether these approaches will work in real practice. For each p, we consider 2-day return, which leads to 377 observations. Let $\mathbf{h}_j = (h_{1t}, \ldots, h_{pj})^\top$, $j = 1, \ldots, 756$, be the daily closing bid prices for each stock at time j. The tth, $t = 1, \ldots, 377$, log-returns for these p stocks are defined as:

$$\mathbf{y}_t = \left(\log \frac{h_{1,2t+1}}{h_{1,2t-1}}, \ldots, \log \frac{h_{p,2t+1}}{h_{p,2t-1}}\right)^\top.$$

We do the following steps:

1. Let $\mathbf{y}_t$, $t = 1, 2, \ldots, 327$, (training data) be the historical data and $\mathbf{y}_{328}$ be the testing data.
2. Suppose the number of factors K is less than $K_{\max} = 40$. We employ Onatski's [27] method to determine the number of factors.
3. Use factor analysis-based approach to derive the estimations $\widehat{\mathbf{B}}$, $\widehat{\mathbf{F}}$ and $\widehat{\Sigma}_u = (\widehat{\sigma}_{ij})$. We suppose there are no correlation of the error terms between two different sectors. Hence, $\widehat{\Sigma}_u^\tau = (\widehat{\sigma}_{ij}^\tau)$ with $\widehat{\sigma}_{ij}^\tau = \widehat{\sigma}_{ij}$ if the ith stock and the jth stock are in the same sector and $\widehat{\sigma}_{ij}^\tau = 0$ if they are in different sector.
4. Estimate $\widehat{\Sigma}_y^{-1}$ by (2.8). Plug-in the sample mean $\bar{\mathbf{y}}$ of the training data in step 1 and $\widehat{\Sigma}_y^{-1}$ to determine the optimal allocation $\widehat{\mathbf{w}}_{f,328}^{(p)}$. We further determine the optimal allocations $\widehat{\mathbf{w}}_{s,328}^{(p)}$, $\widehat{\mathbf{w}}_{p,328}^{(p)}$ and $\widehat{\mathbf{w}}_{b,328}^{(p)}$, by linear shrinkage, plug-in and bootstrap-corrected approaches, respectively.

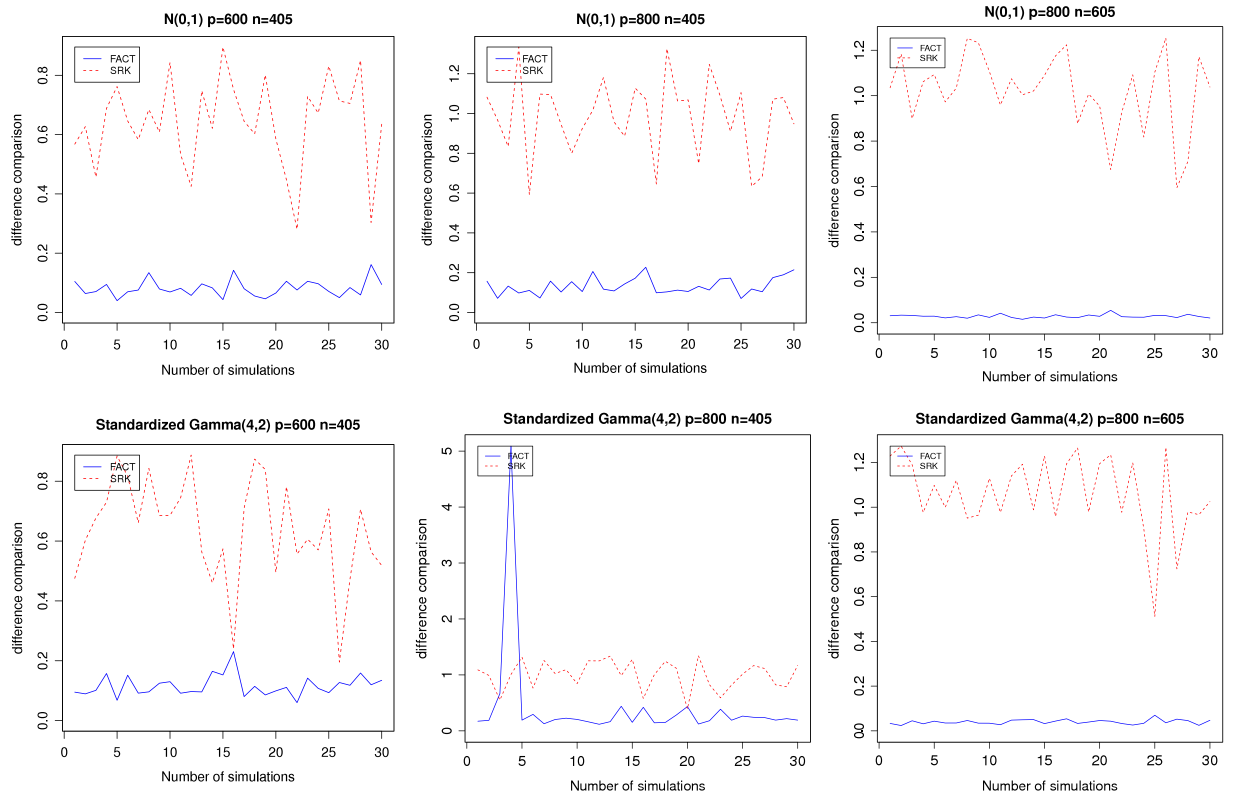

Fig. 4. This figure shows the performance when the data are generated from the factor model (3.1) and the dimension p is bigger than the sample size n. The mean square errors (MSEs) of factor analysis-based approach (blue solid line), and linear shrinkage method (red dotted line) are plotted for each simulation time. The elements of $\mathbf{s}_t$ and $\mathbf{z}_t$, $t = 1, \ldots, n$ are generated from the normal distribution (upper panel) and the standardized Gamma(4, 2) (lower panel). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

The average returns $\bar{R}_l^{(p)}$, $l = f, s, p, b$ for the real data analysis, where $\bar{R}_l^{(p)}$, $l = f, s, p, b$ stand for factor analysis-based (FACT), linear shrinkage (SRK), plug-in (PLN) and bootstrap (BSP) approaches, respectively.

p	$\sigma_0 = 0.1$				$\sigma_0 = 0.5$				$\sigma_0 = 1$			
	FACT ($\bar{R}_f^{(p)}$)	SRK ($\bar{R}_s^{(p)}$)	PLN ($\bar{R}_p^{(p)}$)	BSP ($\bar{R}_b^{(p)}$)	FACT ($\bar{R}_f^{(p)}$)	SRK ($\bar{R}_s^{(p)}$)	PLN ($\bar{R}_p^{(p)}$)	BSP ($\bar{R}_b^{(p)}$)	FACT ($\bar{R}_f^{(p)}$)	SRK ($\bar{R}_s^{(p)}$)	PLN ($\bar{R}_p^{(p)}$)	BSP ($\bar{R}_b^{(p)}$)
100	0.017	0.009	0.026	0.019	0.071	0.034	0.114	0.193	0.139	0.066	0.225	0.252
203	0.022	0.001	−0.03	−0.04	0.097	−0.002	−0.15	0.115	0.192	−0.01	−0.304	−0.843
302	0.034	0.009	–	–	0.167	0.035	–	–	0.332	0.067	–	–
405	0.018	0.004	–	–	0.083	0.009	–	–	0.163	0.015	–	–

5. Applying the optimal allocations $\widehat{\mathbf{w}}_{f,328}^{(p)}$, $\widehat{\mathbf{w}}_{s,328}^{(p)}$, $\widehat{\mathbf{w}}_{p,328}^{(p)}$ and $\widehat{\mathbf{w}}_{b,328}^{(p)}$ to the testing data $\mathbf{y}_{328}$ to obtain the returns $R_{f,328}^{(p)}$, $R_{s,328}^{(p)}$, $R_{p,328}^{(p)}$ and $R_{b,328}^{(p)}$, respectively, where $R_{l,328}^{(p)} = (\widehat{\mathbf{w}}_{l,328}^{(p)})^{\mathsf{T}} \mathbf{y}_{328}$, $l = f, s, p, b$.

Repeating the above steps with training data $\mathbf{y}_t$, $t = 2, \ldots, 328$ and testing data $\mathbf{y}_{329}$, we obtain the returns $R_{l,329}^{(p)}$, $l = f, s, p, b$. Continuing this procedure, we also obtain the returns $R_{l,k}^{(p)}$, $l = f, s, p, b$ for $k = 330, 331, \ldots, 377$. The results of the average returns $\bar{R}_l^{(p)} = \frac{1}{50} \sum_{k=328}^{377} R_{l,k}^{(p)}$, $l = f, s, p, b$, are shown in Table 1.

From Table 1, we observe that when the number of stocks p in a portfolio is not too big compared with the number of observations, bootstrap-corrected and plug-in approach performs better than factor analysis-based approach. However, when p increases, the fact analysis-based approach shows its advantage that it results in bigger average return. Moreover, fact analysis-based approach can be applied to the setting that when $p > n$ or the sample covariance matrix is singular.

Acknowledgments

G.M. Pan was partially supported by the Ministry of Education, Singapore, under a grant #ARC 14/11. S.F. Huang was partially supported by the grant MOST 103-2118-M-390-003 from the Ministry of Science and Technology, Taiwan.

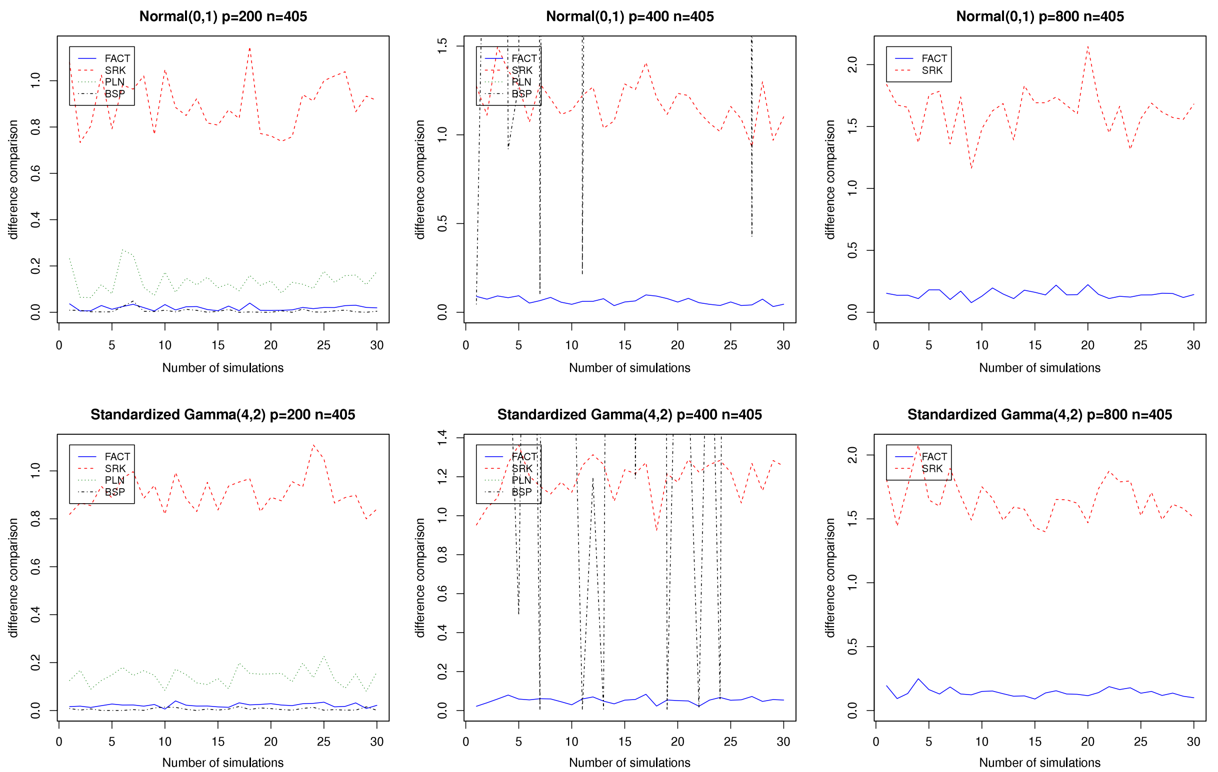

Fig. 5. This figure shows the performance when the data are generated from the multivariate model (3.2). The mean square errors (MSEs) of factor analysis-based approach (blue solid line), linear shrinkage method (red dotted line), plug-in method (green dashed line) and bootstrap-corrected approach (black dashed line) are plotted for each simulation time. Three combinations of $(p, n) = (200, 405), (400, 405), (800, 405)$ are illustrated in the three columns, respectively. The elements of $\mathbf{z}_t,\ t = 1, \ldots, n$ are generated from the normal distribution (upper panel) and the standardized Gamma$(4, 2)$ (lower panel). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix

To begin with, we introduce the following lemma from Petrov [29, p. 254] which will be used to tackle the large deviation problem under the moment conditions.

Lemma A.1. *Let $\{\eta_1, \eta_2, \ldots\}$ be a sequence of independent random variables with $E\eta_i = 0$ and $E\eta_i^2 = \sigma_i^2 > 0$ for all i. Define $S_n = \sum_{i=1}^n \eta_i$ and $V_n = \sum_{i=1}^n \sigma_i^2$. Suppose*

$$\liminf_{n\to\infty}\left(\frac{V_n}{n}\right) > 0 \quad \text{and} \quad \liminf_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^n E|\eta_i|^k\right) < \infty$$

for some $k > 2$. Let $\Phi(x) = \int_{-\infty}^x (2\pi)^{1/2} e^{-t^2/2} dt$. Then

$$\frac{P\left(\frac{S_n}{\sqrt{V_n}} > x\right)}{1 - \Phi(x)} \to 1$$

uniformly on $[0, c\sqrt{\log n}]$ for any $c \in (0, \sqrt{k-2})$ as $n \to \infty$.

Lemma A.2. *Under Assumptions A–D, we have*

(i)

$$P\left(\max_{i\le p}\left|\frac{1}{n}\sum_{t=1}^n u_{it}\right| \le \sqrt{\frac{2c_1 \log p}{n}}\right) = 1 - o(1),$$

(ii)

$$P\left(\max_{i\le p}\left|\frac{1}{n}\sum_{t=1}^n u_{it}^2 - \sigma_{ii}\right| \le \sqrt{\frac{2c_1 \log p}{n}}\right) = 1 - o(1),$$

(iii)

$$P\Big(\max_{i,j\le p}\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it}u_{jt}-\sigma_{ij}\Big|\le\sqrt{\frac{4c_1\log p}{n}}\Big)=1-o(1),$$

(iv)

$$P\Big(\max_{k\le K,}\Big|\frac{1}{n}\sum_{t=1}^{n}f_{kt}\Big|\le\sqrt{\frac{c_3 m_n}{n}}\Big)=1-o(1),$$

(v)

$$P\Big(\max_{k\le K,i\le p}\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it}f_{kt}\Big|\le\sqrt{\frac{2c_1c_3\log p}{n}}\Big)=1-o(1),$$

where c_1, c_3 are defined in Assumptions B and D, respectively and m_n tends to infinity with $m_n/n\to 0$.

Proof of Lemma A.2. In view of Assumption A, we have $\sqrt{2\log p}\in(0,\sqrt{2\alpha\log n})$. Then by Lemma A.1, under the condition $E|u_{it}|^{(2\alpha+2)+\delta}$, we have for n large

$$P\Big(\Big|\frac{\frac{1}{n}\sum_{t=1}^{n}u_{it}}{\sqrt{\sigma_{ii}/n}}\Big|>\sqrt{2\log p}\Big)=2\Big(1-\Phi(\sqrt{2\log p})\Big)\approx\frac{2}{p\sqrt{4\pi\log p}},$$

where we use the fact that

$$1-\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt\approx\frac{e^{-x^2/2}}{\sqrt{2\pi}x}. \tag{A.1}$$

Noting that $\max_{i\le p}\sigma_{ii}\le c_1$, we have

$$\begin{aligned}P\Big(\max_{i\le p}\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it}\Big|>\sqrt{\frac{2c_1\log p}{n}}\Big)&\le\sum_{i=1}^{p}P\Big(\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it}\Big|>\sqrt{\frac{2c_1\log p}{n}}\Big)\\&\le\sum_{i=1}^{p}P\Big(\Big|\frac{\frac{1}{n}\sum_{t=1}^{n}u_{it}}{\sqrt{\sigma_{ii}/n}}\Big|>\sqrt{2\log p}\Big)\\&\le p\cdot\frac{1}{p\sqrt{4\pi\log p}}=o(1).\end{aligned}$$

Therefore Lemma A.2(i) is proved. Estimates (ii)–(v) can be similarly proved by Lemma A.1 and (A.1). Indeed, to prove (i), (v), we need the condition $E|u_{it}|^{(2\alpha+2)+\delta}<\infty$ for $i\le p$ and some $\delta>0$. To prove (ii), we need $E|u_{it}|^{2(2\alpha+2)+\delta}<\infty$ for $i\le p$ and for (iii), the condition $E|u_{it}u_{jt}|^{(4\alpha+2)+\delta}<\infty$ for $i\le p$ is needed. □

Proof of Lemma 2.1(i). Since $\frac{1}{p}\mathbf{B}\mathbf{B}^\top$, Σ_f are $K\times K$ matrices with fixed K, their spectral norms are bounded from above by Assumptions C and D. We assume there exists a constant c_3, such that

$$\Big\|\frac{1}{p}\mathbf{B}\mathbf{B}^\top\Big\|,\ \|\Sigma_f\|,\ \max_{i,j\le K}|\rho_{ij}|\le c_3, \tag{A.2}$$

where ρ_{ij} is defined in Assumption D.

For the simplification of the notation, let

$$\mathbf{F}_*=\mathbf{F}-\bar{\mathbf{f}}\mathbf{1}_n^\top=(f_{kt*}),\qquad \mathbf{U}_*=\mathbf{U}-\bar{\mathbf{u}}\mathbf{1}_n^\top=(u_{it*}),\qquad \mathbf{Y}_*=\mathbf{Y}-\bar{\mathbf{y}}\mathbf{1}_n^\top=\mathbf{B}^\top\mathbf{F}_*+\mathbf{U}_*, \tag{A.3}$$

where $\bar{\mathbf{f}}=\frac{1}{n}\sum_{t=1}^n\mathbf{f}_t$ and $\bar{\mathbf{u}}=\frac{1}{n}\sum_{t=1}^n\mathbf{u}_t=(\bar{u}_1,\ldots,\bar{u}_p)^\top$. Note $\bar{\mathbf{y}}=\boldsymbol{\mu}+\mathbf{B}^\top\bar{\mathbf{f}}+\bar{\mathbf{u}}$. Suppose $\Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_K)$ where $\lambda_1,\ldots,\lambda_K$ are the first K eigenvalues of $\frac{1}{np}\mathbf{Y}_*\mathbf{Y}_*^\top$. Thus, $\widehat{\mathbf{B}}^\top=\frac{1}{np}\mathbf{Y}_*\mathbf{Y}_*^\top\widehat{\mathbf{B}}^\top\Lambda^{-1}$. We first show $\lambda_K\ge c_0^2/2>0$, which ensures that Λ^{-1} is well defined. Let $s_k(\mathbf{M})$, $k=1,2,\ldots$ denote the singular values of the matrix $\mathbf{M}$ in the decreasing order. By Weyl's inequality

$$\begin{aligned}\sqrt{\lambda_K}&=\lambda_K^{1/2}\Big(\frac{1}{np}(\mathbf{B}^\top\mathbf{F}_*+\mathbf{U}_*)(\mathbf{B}^\top\mathbf{F}_*+\mathbf{U}_*)^\top\Big)=s_K\Big(\frac{1}{\sqrt{np}}(\mathbf{B}^\top\mathbf{F}_*+\mathbf{U}_*)\Big)\\&\ge s_K(\mathbf{B}^\top\mathbf{F}_*/\sqrt{np})-\|\mathbf{U}_*/\sqrt{np}\|\\&=\lambda_{\min}^{1/2}\Big(\frac{1}{p}\mathbf{B}\mathbf{B}^\top\cdot\frac{1}{n}\mathbf{F}_*\mathbf{F}_*^\top\Big)-\lambda_{\max}^{1/2}\Big(\frac{1}{np}\mathbf{U}_*\mathbf{U}_*^\top\Big).\end{aligned} \tag{A.4}$$

Since K is fixed, via Assumption D, we have $\frac{1}{n}\mathbf{F}_*\mathbf{F}_*^\top \xrightarrow{\text{i.p.}} \Sigma_f$ in probability. Hence by Assumption D and (A.2), as $n \to \infty$,

$$\begin{aligned}
&\lambda_{\min}\left(\frac{1}{n}\mathbf{F}_*\mathbf{F}_*^\top\right) \xrightarrow{\text{i.p.}} \lambda_{\min}(\Sigma_f) \ge c_0, \\
&\lambda_{\max}\left(\frac{1}{n}\mathbf{F}_*\mathbf{F}_*^\top\right) \xrightarrow{\text{i.p.}} \lambda_{\max}(\Sigma_f) \le c_3.
\end{aligned} \tag{A.5}$$

By Assumption C and the fact that $\lambda_{\min}(AB) \ge \lambda_{\min}(A)\cdot\lambda_{\min}(B)$ for any conformable positive matrix A, B, we have

$$\lambda_{\min}\left(\frac{1}{p}\mathbf{B}\mathbf{B}^\top \cdot \frac{1}{n}\mathbf{F}_*\mathbf{F}_*^\top\right) \ge \lambda_{\min}\left(\frac{1}{n}\mathbf{F}_*\mathbf{F}_*^\top\right)\cdot\lambda_{\min}\left(\frac{1}{p}\mathbf{B}\mathbf{B}^\top\right) \ge c_0^2, \tag{A.6}$$

in probability as $n \to \infty$. On the other hand, when p is proportional to n, $\lambda_{\max}\left(\frac{1}{n}\mathbf{U}_*\mathbf{U}_*^\top\right) \le C$ with probability one for some constant C by Pan [28]. When $p/n \to \infty$, $\lambda_{\max}\left(\frac{1}{p}\mathbf{U}\mathbf{U}^\top\right) \le C$ with probability one by Chen and Pan [5]. Thus

$$\lambda_{\max}\left(\frac{1}{\max(n,p)}\mathbf{U}_*\mathbf{U}_*^\top\right) \le C \tag{A.7}$$

in probability for some constant C. Then we have $\lambda_{\max}\left(\frac{1}{np}\mathbf{U}_*\mathbf{U}_*^\top\right) \xrightarrow{\text{i.p.}} 0$. This, together with (A.4) and (A.6), implies $\lambda_K \ge c_0^2 - 2\sqrt{c_0^2\cdot o_p(1)} \ge c_0^2/2$ in probability. Hence, Λ^{-1} exists and its spectral norm is bounded by $2/c_0^2$ in probability.

Let $\Upsilon = \frac{1}{np}\mathbf{F}_*\mathbf{F}_*^\top\mathbf{B}\widehat{\mathbf{B}}^\top\Lambda^{-1}$. We then have

$$\begin{aligned}
\widehat{\mathbf{B}}^\top - \mathbf{B}^\top\Upsilon &= \frac{1}{np}\mathbf{Y}_*\mathbf{Y}_*^\top\widehat{\mathbf{B}}\Lambda^{-1} - \mathbf{B}^\top\Upsilon \\
&= \frac{1}{np}(\mathbf{B}^\top\mathbf{F}_* + \mathbf{U}_*)(\mathbf{B}^\top\mathbf{F}_* + \mathbf{U}_*)^\top\widehat{\mathbf{B}}^\top\Lambda^{-1} - \mathbf{B}^\top\Upsilon \\
&= \frac{1}{np}\mathbf{B}^\top\mathbf{F}_*\mathbf{U}_*^\top\widehat{\mathbf{B}}^\top\Lambda^{-1} + \frac{1}{np}\mathbf{U}_*\mathbf{F}_*^\top\mathbf{B}\widehat{\mathbf{B}}^\top\Lambda^{-1} + \frac{1}{np}\mathbf{U}_*\mathbf{U}_*^\top\widehat{\mathbf{B}}^\top\Lambda^{-1}.
\end{aligned} \tag{A.8}$$

From (2.6), (A.3) and the fact that $\frac{1}{p}\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top = \mathbf{I}_K$, write

$$\begin{aligned}
\Upsilon\widehat{\mathbf{F}} &= \frac{1}{p}\Upsilon\widehat{\mathbf{B}}\mathbf{Y}_* = \frac{1}{p}\Upsilon\widehat{\mathbf{B}}\mathbf{B}^\top\mathbf{F}_* + \frac{1}{p}\Upsilon\widehat{\mathbf{B}}\mathbf{U}_* \\
&= \frac{1}{p}\Upsilon\widehat{\mathbf{B}}\widehat{\mathbf{B}}^\top\Upsilon^{-1}\mathbf{F}_* + \frac{1}{p}\Upsilon\widehat{\mathbf{B}}(\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}}^\top)\Upsilon^{-1}\mathbf{F}_* + \frac{1}{p}\Upsilon\widehat{\mathbf{B}}\mathbf{U}_* \\
&= \mathbf{F}_* + \frac{1}{p}\Upsilon\widehat{\mathbf{B}}(\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}}^\top)\Upsilon^{-1}\mathbf{F}_* + \frac{1}{p}\Upsilon\widehat{\mathbf{B}}\mathbf{U}_*.
\end{aligned}$$

Hence,

$$\Upsilon\widehat{\mathbf{F}} - \mathbf{F}_* = \frac{1}{p}\Upsilon\widehat{\mathbf{B}}(\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}}^\top)\Upsilon^{-1}\mathbf{F}_* + \frac{1}{p}\Upsilon\widehat{\mathbf{B}}\mathbf{U}_*. \tag{A.9}$$

It follows from (2.6), (A.3) and (A.9) that

$$\begin{aligned}
\mathbf{U}_* - \widehat{\mathbf{U}} &= \widehat{\mathbf{B}}^\top\widehat{\mathbf{F}} - \mathbf{B}^\top\mathbf{F}_* = (\widehat{\mathbf{B}}^\top - \mathbf{B}^\top\Upsilon)\widehat{\mathbf{F}} + \mathbf{B}^\top(\Upsilon\widehat{\mathbf{F}} - \mathbf{F}_*) \\
&= (\widehat{\mathbf{B}}^\top - \mathbf{B}^\top\Upsilon)\widehat{\mathbf{F}} + \frac{1}{p}\mathbf{B}^\top\Upsilon\widehat{\mathbf{B}}(\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}})\Upsilon^{-1}\mathbf{F}_* + \frac{1}{p}\mathbf{B}^\top\Upsilon\widehat{\mathbf{B}}\mathbf{U}_*.
\end{aligned} \tag{A.10}$$

For sufficiently large but fixed C, let the event

$$A_0 = \left\{\left\|\frac{1}{p}\mathbf{B}\mathbf{B}^\top\right\|, \left\|\frac{1}{\max(n,p)}\mathbf{U}_*\mathbf{U}_*^\top\right\|, \left\|\frac{1}{n}\mathbf{F}_*\mathbf{F}_*^\top\right\|, \left\|\frac{1}{n}\widehat{\mathbf{F}}\widehat{\mathbf{F}}^\top\right\|, \|\Upsilon\|, \|\Upsilon^{-1}\|, \|\Lambda\|, \|\Lambda^{-1}\| \le C\right\}.$$

We further define the following events:

$$A_1 = \left\{\max_{i\le p}\left|\frac{1}{n}\sum_{t=1}^n u_{it}\right| \le \sqrt{\frac{2c_1\log p}{n}}\right\}, \qquad A_2 = \left\{\max_{i\le p}\left|\frac{1}{n}\sum_{t=1}^n u_{it}^2 - \sigma_{ii}\right| \le \sqrt{\frac{2c_1\log p}{n}}\right\}$$

$$A_3 = \left\{ \max_{i,j\le p} \left| \frac{1}{n}\sum_{t=1}^{n} u_{it}u_{jt} - \sigma_{ij} \right| \le \sqrt{\frac{4c_1 \log p}{n}} \right\}, \qquad A_4 = \left\{ \max_{k\le K,} \left| \frac{1}{n}\sum_{t=1}^{n} f_{kt} \right| \le \sqrt{\frac{c_3 m_n}{n}} \right\}$$

$$A_5 = \left\{ \max_{k\le K, i\le p} \left| \frac{1}{n}\sum_{t=1}^{n} u_{it} f_{kt} \right| \le \sqrt{\frac{2c_1 c_3 \log p}{n}} \right\}.$$

From Lemma A.2, we have $P(A_i) = 1 - o(1),\ i = 1, 2, \ldots, 5$. We claim that $P(A_0) = 1 - o(1)$. Indeed, by Assumption C, (A.4), (A.5) and (A.7)

$$\left\| \frac{1}{p}\mathbf{B}\mathbf{B}^\top \right\| \le c_2, \qquad \left\| \frac{1}{\max(n,p)} \mathbf{U}_*\mathbf{U}_*^\top \right\| \le C, \qquad \left\| \frac{1}{n}\mathbf{F}_*\mathbf{F}_*^\top \right\| \le c_3, \qquad \|\Lambda^{-1}\| \le 2/c_0^2,$$

$$\|\Lambda\| = \lambda_1 \le \frac{2}{np}\|\mathbf{B}^\top\mathbf{F}_*\mathbf{F}_*^\top\mathbf{B}\| + \frac{2}{np}\|\mathbf{U}_*\mathbf{U}_*^\top\| \le C,$$

$$\|\Upsilon\| \le \frac{1}{np}\|\mathbf{B}\widehat{\mathbf{B}}^\top\mathbf{F}_*\mathbf{F}_*^\top\|\,\|\Lambda^{-1}\| \le C,$$

$$\lambda_{\min}(\Upsilon) \ge \lambda_{\min}\left(\frac{1}{p}\mathbf{B}\mathbf{B}^\top\right)\cdot\lambda_{\min}\left(\frac{1}{p}\mathbf{F}_*\mathbf{F}_*^\top\right)\cdot\lambda_{\min}(\Lambda) \ge C,$$

in probability. By (2.6), we obtain

$$\left\| \frac{1}{n}\widehat{\mathbf{F}\mathbf{F}}^\top \right\| \le \left\| \frac{1}{p}\widehat{\mathbf{B}\mathbf{B}}^\top \right\| \cdot \left\| \frac{1}{np}\mathbf{Y}_*\mathbf{Y}_*^\top \right\| \le C,$$

in probability, where $\widehat{\mathbf{F}}$ is defined in (2.6). Hence, $P(A_0) = 1 - o(1)$.

Throughout the remaining proof, we assume the events $A_i,\ i = 0, 1, \ldots, 5$ happen. Let $\mathbf{e}_i$ be the p-dimensional unit vector with the ith component being 1 and others being 0. Then by (A.10), we have

$$\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it} - u_{it*})^2 = \frac{1}{n}\mathbf{e}_i^\top(\mathbf{U}_* - \widehat{\mathbf{U}})(\mathbf{U}_* - \widehat{\mathbf{U}})^\top\mathbf{e}_i \le C(J_1 + J_2 + J_3),$$

where

$$J_1 = \frac{1}{n}\mathbf{e}_i^\top(\widehat{\mathbf{B}}^\top - \mathbf{B}^\top\Upsilon)\widehat{\mathbf{F}\mathbf{F}}^\top(\widehat{\mathbf{B}}^\top - \mathbf{B}^\top\Upsilon)^\top\mathbf{e}_i, \qquad J_3 = \frac{1}{np^2}\mathbf{b}_i^\top\Upsilon\widehat{\mathbf{B}}\mathbf{U}_*\mathbf{U}_*^\top\widehat{\mathbf{B}}^\top\Upsilon^\top\mathbf{b}_i,$$

$$J_2 = \frac{1}{np^2}\mathbf{b}_i^\top\Upsilon\widehat{\mathbf{B}}(\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}})\Upsilon^{-1}\mathbf{F}_*\mathbf{F}_*^\top(\Upsilon^{-1})^\top\Upsilon\widehat{\mathbf{B}}(\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}})^\top\mathbf{b}_i.$$

When the event A_0 happens, it follows from (A.8) that

$$\|\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}}^\top\| \le C\Big(1 + \sqrt{p/n} + \max(p,n)/n\sqrt{p}\Big) \le C\Big(1 + \sqrt{\frac{p}{n}}\Big), \tag{A.11}$$

and

$$\begin{aligned}\|\mathbf{e}_i^\top(\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}}^\top)\| &= \left\| -\frac{1}{np}\mathbf{b}_i^\top\mathbf{F}_*\mathbf{U}_*^\top\widehat{\mathbf{B}}^\top\Lambda^{-1} - \frac{1}{np}\left(\sum_{t=1}^{n}u_{it*}\mathbf{f}_{t*}^\top\right)\mathbf{B}^\top\widehat{\mathbf{B}}^\top\Lambda^{-1} + \frac{1}{np}(\mathbf{e}_i^\top\mathbf{U}_*)\mathbf{U}_*^\top\widehat{\mathbf{B}}^\top\Lambda^{-1} \right\| \\ &\le \sqrt{\mathbf{b}_i^\top\mathbf{b}_i}\cdot\max\Big(\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{n}}\Big) + \frac{1}{\sqrt{n}}\cdot\left\| \frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{it*}\mathbf{f}_t^\top \right\| + \Big(\frac{1}{n}\sum_{t=1}^{n}u_{it*}^2\Big)^{1/2}\cdot\max\Big(\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{n}}\Big).\end{aligned} \tag{A.12}$$

Consider J_1 now. By (A.12), we have

$$|J_1| \le \|\mathbf{e}_i^\top(\mathbf{B}^\top\Upsilon - \widehat{\mathbf{B}}^\top)\|^2 \le \frac{3}{\min(p,n)}\max_{i\le p}\Big(\mathbf{b}_i^\top\mathbf{b}_i + \frac{3}{n}\sum_{t=1}^{n}u_{it*}^2\Big) + \frac{1}{n}\max_{i\le p}\Big[\Big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{it*}\mathbf{f}_t^\top\Big)\Big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{it*}\mathbf{f}_t\Big)\Big].$$

When the events A_1, A_2 happen, we have

$$\max_{i\le p}\left| \frac{1}{n}\sum_{t=1}^{n}u_{it*}^2 \right| \le 3\max_{i\le p}\left| \frac{1}{n}\sum_{t=1}^{n}u_{it}^2 - \sigma_{ii} \right| + 3\max_{i\le p}|\sigma_{ii}| + 3\max_{i\le p}\left| \frac{1}{n}\sum_{t=1}^{n}u_{it} \right| \le \sqrt{\frac{2c_1\log p}{n}} + 2c_1 \le 3c_1.$$

When the events A_1, A_4, A_5 occur, we have

$$\begin{aligned}\max_{i\le p}\Big[\Big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{it*}\mathbf{f}_t^{\mathsf{T}}\Big)\Big(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{it*}\mathbf{f}_t\Big)\Big] &\le K\max_{k\le K,i\le p}\Big|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{it*}f_{kt}\Big|^2\\ &\le 2K\max_{k\le K,i\le p}\Big|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}u_{it}f_{kt}\Big|^2+2K^2\max_{k\le K}\Big|\frac{1}{\sqrt{n}}\sum_{t=1}^{n}f_{kt}\Big|^2\cdot\max_{i\le p}\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it}\Big|^2\\ &\le 4Kc_1c_3\log p+4K^2c_1c_3m_n\log p/n\\ &\le 5Kc_1c_3\log p.\end{aligned}$$

Hence, via Assumption B,

$$|J_1|\le\frac{3c_1+Kc_2}{\min(n,p)}+\frac{5K^2c_1c_3\log p}{n}\le C\Big(\frac{1}{p}+\frac{\log p}{n}\Big).\tag{A.13}$$

Consider J_2. By (A.11) and Assumption C,

$$|J_2|\le\max_{i\le p}(\mathbf{b}_i^{\mathsf{T}}\mathbf{b}_i)\cdot\frac{1}{p}\|\mathbf{B}^{\mathsf{T}}\Upsilon-\widehat{\mathbf{B}}^{\mathsf{T}}\|^2\le\frac{c_2}{\min(p,n)}.\tag{A.14}$$

Consider J_3. When A_0 happens,

$$|J_3|\le\frac{c_2}{\min(p,n)}.\tag{A.15}$$

Therefore, we conclude from (A.13)–(A.15) that

$$\max_{i\le p}\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}-u_{it*})^2\le C\Big(\frac{1}{p}+\frac{\log p}{n}\Big).$$

The proof of Lemma 2.1(i) is completed.

Proof of Lemma 2.1(ii). When the events A_1, A_2 happen, we have

$$\begin{aligned}\max_{i\le p}\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it*}^2\Big| &\le 2\max_{i\le p}\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it}^2\Big|+2\max_{i\le p}\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it}\Big|^2\\ &\le 2\max_{i\le p}\sigma_{ii}+o(1)\le C.\end{aligned}$$

By Holder's inequality and Lemma 2.1(i), we have

$$\begin{aligned}\max_{i,j}\Big|\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}\widehat{u}_{jt}-u_{it*}u_{jt*})\Big| &\le \max_{i\le p}\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}-u_{it*})^2+\max_{i\le p}\sqrt{\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}-u_{it*})^2}\cdot\max_{i\le p}\sqrt{\frac{1}{n}\sum_{t=1}^{n}u_{it*}^2}\\ &\le C\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big).\quad\square\end{aligned}$$

Lemma A.3. *Under assumptions of Theorem 2.1,*

$$P\Big(c_l\le\min_{i,j\le p}\widehat{\theta}_{ij}\le\max_{i,j\le p}\widehat{\theta}_{ij}\le c_u\Big)=1-o(1),$$

where $c_l=c_0/5$, $c_u=17c_1$ and c_0, c_1 are given in Assumption B.

Proof of Lemma A.3. Define the events

$$V_1=\Big\{\max_{i\le p}\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}-u_{it*})^2\le C\Big(\frac{1}{p}+\frac{\log p}{n}\Big)\Big\},$$

$$V_2=\Big\{\max_{i,j\le p}\Big|\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}\widehat{u}_{jt}-u_{it*}u_{jt*})\Big|\le C\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)\Big\}.$$

It follows from Lemma 2.1 that $P(V_i) = 1 - o(1)$. We next suppose the events $A_i,\ i = 0, \dots, 5$ and $V_1,\ V_2$ happen. By the inequality $(a_1 + \cdots + a_m)^2 \le m(a_1^2 + \cdots + a_m^2)$, we decompose $\widehat{\theta}_{ij}$ as

$$\begin{aligned}
\widehat{\theta}_{ij} &= \frac{1}{n}\sum_{t=1}^{n}\left[(\widehat{u}_{it}\widehat{u}_{jt} - u_{it*}u_{jt*}) + (u_{it*}u_{jt*} - \sigma_{it}) + \left(\sigma_{ij} - \frac{1}{n}\sum_{t=1}^{n} u_{it*}u_{jt*}\right) + \left(\frac{1}{n}\sum_{t=1}^{n} u_{it*}u_{jt*} - \widehat{\sigma}_{ij}\right)\right]^2 \\
&\le \frac{4}{n}\sum_{t=1}^{n}(\widehat{u}_{it}\widehat{u}_{jt} - u_{it*}u_{jt*})^2 + \frac{4}{n}\sum_{t=1}^{n}(u_{it*}u_{jt*} - \sigma_{it})^2 \\
&\quad + 4\left|\frac{1}{n}\sum_{t=1}^{n} u_{it*}u_{jt*} - \sigma_{ij}\right|^2 + 4\left|\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}\widehat{u}_{jt} - u_{it*}u_{jt*})\right|^2 \\
&\triangleq \iota_1 + \iota_2 + \iota_3 + \iota_4,
\end{aligned}$$

where $\iota_i,\ i = 1, 2, 3, 4$ are defined in an obvious way.

We next show $\max_{i,j\le p}\left(\iota_1, \iota_3, \iota_4\right)$ will converge to zero and $\max_{i,j\le p}\iota_2$ will be smaller than a positive constant in probability. Consider ι_1. As in the proof of Lemma A.2, we have $\max_{i\le p}\left|\frac{1}{n}\sum_{t=1}^{n} u_{it*}^4 - Eu_{it}^4\right| \le C\sqrt{\frac{\log p}{n}}$ under the condition $E|u_{it}|^{4(2\alpha+2)+\delta} < \infty$. This, together with the fact that $\max_{i\le p}\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it} - u_{it*})^4 \le n\max_{i\le p}\left(\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it} - u_{it*})^2\right)^2 \le C\left(\frac{n}{p^2} + \frac{\log^2 p}{n}\right)$, implies

$$\begin{aligned}
\iota_1 &= \frac{4}{n}\sum_{t=1}^{n}\Big[(\widehat{u}_{it} - u_{it*})(\widehat{u}_{jt} - u_{jt*}) + u_{it*}(\widehat{u}_{jt} - u_{jt*}) + u_{jt*}(\widehat{u}_{it} - u_{it*})\Big]^2 \\
&\le \max_{i,j\le p}\frac{12}{n}\sum_{t=1}^{n}(\widehat{u}_{it} - u_{it*})^2(\widehat{u}_{jt} - u_{jt*})^2 + \max_{i,j\le p}\frac{24}{n}\sum_{t=1}^{n} u_{it*}^2(\widehat{u}_{jt} - u_{jt*})^2 \\
&\le \max_{i\le p}\frac{12}{n}\sum_{t=1}^{n}(\widehat{u}_{it} - u_{it*})^4 + 24\sqrt{\max_{i\le p}\frac{1}{n}\sum_{t=1}^{n} u_{it*}^4 \cdot \max_{i\le p}\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{jt} - u_{jt*})^4} \\
&\le C\sqrt{\frac{n}{p^2} + \frac{\log^2 p}{n}}.
\end{aligned}$$

Consider ι_2. Since $u_{it*} = u_{it} - \bar{u}_i$, when the events A_1, A_2 happen,

$$\begin{aligned}
\iota_2 &= \frac{4}{n}\sum_{t=1}^{n}\Big[(u_{it}u_{jt} - \sigma_{ij}) - \bar{u}_i u_{jt} - \bar{u}_j u_{it} + \bar{u}_i\bar{u}_j\Big]^2 \\
&\le \max_{i,j\le p}\frac{16}{n}\sum_{t=1}^{n}\Big((u_{it}u_{jt} - \sigma_{ij})^2 - \operatorname{var}(u_{it}u_{jt})\Big) + 16\operatorname{var}(u_{it}u_{jt}) \\
&\quad + 32\max_{i\le p}\left|\frac{1}{n}\sum_{t=1}^{n} u_{it}\right|^2 \cdot \max_{i\le p}\frac{1}{n}\sum_{t=1}^{n} u_{it}^2 + 16\max_{i\le p}\left|\frac{1}{n}\sum_{t=1}^{n} u_{it}\right|^4 \\
&\le 16\operatorname{var}(u_{it}u_{jt}) + C\sqrt{\frac{\log p}{n}},
\end{aligned}$$

where we use the fact that $\max_{i,j\le p}\left|\frac{16}{n}\sum_{t=1}^{n}\Big((u_{it}u_{jt} - \sigma_{ij})^2 - \operatorname{var}(u_{it}u_{jt})\Big)\right| \le C\sqrt{\frac{\log p}{n}}$ in probability under the condition $E|u_{it}u_{jt}|^{2(4\alpha+2)+\delta} < \infty$, which can be similarly proved as in Lemma A.2. When the events A_1, A_2, V_2 happen, without any difficulty, we have

$$\max_{i,j\le p}\iota_3 \le C\sqrt{\frac{\log p}{n}} \quad\text{and}\quad \max_{i,j\le p}\iota_4 \le C\left(\frac{1}{p} + \sqrt{\frac{\log p}{n}}\right).$$

Therefore, from Assumption B, $\max_{i,j\le p}\widehat{\theta}_{ij} \le \max_{i,j\le p} 16\operatorname{var}(u_{it}u_{jt}) + C\sqrt{\frac{n}{p^2} + \frac{\log^2 p}{n}} \le 17c_1$ in probability, where we use the assumption that $n/p^2 \to 0$ in Theorem 2.1.

On the other hand, since

$$\frac{1}{n}\sum_{t=1}^{n}(u_{it*}u_{jt*}-\sigma_{it})^2 = \frac{1}{n}\sum_{t=1}^{n}\left[(\widehat{u}_{it}\widehat{u}_{jt}-\widehat{\sigma}_{ij})-(\widehat{u}_{it}\widehat{u}_{jt}-u_{it*}u_{jt*})-\left(\sigma_{ij}-\frac{1}{n}\sum_{t=1}^{n}u_{it*}u_{jt*}\right)\right.$$
$$\left.-\left(\frac{1}{n}\sum_{t=1}^{n}u_{it*}u_{jt*}-\widehat{\sigma}_{ij}\right)\right]^2$$
$$\leq 4\widehat{\theta}_{ij}+\iota_1+\iota_3+\iota_4,$$

we have

$$\min_{i,j\leq p}\widehat{\theta}_{ij} \geq \min_{i,j\leq p}\frac{1}{4n}\sum_{t=1}^{n}\Big[(u_{it*}u_{jt*}-\sigma_{it})^2-\mathrm{var}(u_{it}u_{jt})\Big]+\frac{1}{4}\mathrm{var}(u_{it}u_{jt})-C\sqrt{\frac{n}{p^2}+\frac{\log^2 p}{n}}$$
$$\geq \frac{1}{5}\mathrm{var}(u_{it}u_{jt}),$$

where we also use the assumption that $n/p^2 \to 0$. We remind the readers that we only use the assumption n/p^2 here. Hence, $\min_{i,j\leq p}\widehat{\theta}_{ij}\geq c_0/5$ in probability. Lemma A.3 is then proved by letting

$$c_l = c_0/5 \quad\text{and}\quad c_u = 17c_1. \quad\square$$

Proof of Theorem 2.1. For the spectral norm, we have

$$\|\widehat{\Sigma}_u^{\tau}-\Sigma_u\| \leq \max_{i\leq p}\sum_{j=1}^{p}\Big|\widehat{\sigma}_{ij}I(|\widehat{\sigma}_{ij}\geq\omega_n\sqrt{\widehat{\theta}_{ij}}|)-\sigma_{ij}\Big|.$$

We define the events

$$V_3=\left\{\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|\leq C\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)\right\},\qquad V_4=\left\{c_l\leq\min_{i,j\leq p}\widehat{\theta}_{ij}\leq\max_{i,j\leq p}\widehat{\theta}_{ij}\leq c_u\right\}.$$

Under the events A_1, A_3, V_2, we have

$$\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}| \leq \max_{i,j\leq p}\Big|\frac{1}{n}\sum_{t=1}^{n}(\widehat{u}_{it}\widehat{u}_{jt}-u_{it*}u_{jt*})\Big|+\max_{i,j\leq p}\Big|\frac{1}{n}\sum_{t=1}^{n}(u_{it}u_{jt}-\sigma_{ij})\Big|+\max_{i\leq p}\Big|\frac{1}{n}\sum_{t=1}^{n}u_{it}\Big|^2$$
$$\leq C\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)+\sqrt{\frac{4c_1\log p}{n}}+\frac{2c_1\log p}{n}$$
$$\leq C\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big). \tag{A.16}$$

Hence $P(V_3)=1-o(1)$. From Lemma A.3, $P(V_4)=1-o(1)$. We suppose the events $A_i,\ i=0,1,\ldots,5$ and $V_i,\ i=1,\ldots,4$ happen. If we choose $\omega_n=\frac{2C}{\sqrt{c_l}}\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)$, then event $|\widehat{\sigma}_{ij}|\geq\omega_n\widehat{\theta}_{ij}^{1/2}$ implies $|\sigma_{ij}|\geq|\widehat{\sigma}_{ij}|-\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|\geq C\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)$. And the event $|\widehat{\sigma}_{ij}|<\omega_n\widehat{\theta}_{ij}^{1/2}$ implies $|\sigma_{ij}|\leq|\widehat{\sigma}_{ij}|+\max_{i,j\leq p}|\widehat{\sigma}_{ij}-\sigma_{ij}|\leq C\Big(2\sqrt{\frac{c_u}{c_l}}+1\Big)\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)$. Therefore, by (A.16) and $\Sigma_u\in\mathbb{U}_p\Big(C,\kappa(p,n)\Big)$, we have, uniformly in $i\leq p$,

$$\|\widehat{\Sigma}_u^{\tau}-\Sigma_u\| \leq \sum_{j=1}^{p}\Big|\widehat{\sigma}_{ij}I(|\widehat{\sigma}_{ij}\geq\omega_n\sqrt{\widehat{\theta}_{ij}}|)-\sigma_{ij}\Big|$$
$$\leq \sum_{j=1}^{p}|\widehat{\sigma}_{ij}-\sigma_{ij}|I(\widehat{\sigma}_{ij}\geq\omega_n\widehat{\theta}_{ij}^{1/2})+\sum_{j=1}^{p}|\sigma_{ij}|I(\widehat{\sigma}_{ij}<\omega_n\widehat{\theta}_{ij}^{1/2})$$
$$\leq \sum_{j=1}^{p}|\widehat{\sigma}_{ij}-\sigma_{ij}|I\Big(|\sigma_{ij}|>C\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)\Big)+\sum_{j=1}^{p}|\sigma_{ij}|I\Big(|\sigma_{ij}|\leq C\Big(2\sqrt{\frac{c_u}{c_l}}+1\Big)\Big(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\Big)\Big)$$
$$\leq C\left(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\right)\cdot\max_{i\leq p}\sum_{j=1}^{p}I\left(|\sigma_{ij}|>C\left(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\right)\right)$$

$$+\max_{i\le p}\sum_{j=1}^{p}|\sigma_{ij}|I\left(|\sigma_{ij}|\le C\left(\frac{1}{\sqrt{p}}+\sqrt{\frac{\log p}{n}}\right)\right)$$
$$\le C\kappa(p,n).$$

Hence, Theorem 2.1 is proved.

Proof of Theorem 2.2. Let the event $V_5=\left\{\|\widehat{\Sigma}_u^{\tau}-\Sigma_u\|\le C\kappa(p,n)\right\}$. Then by Theorem 2.1, we have $P(V_5)=1-o(1)$. We suppose the events $A_i,\ i=0,\dots,5$ and $V_i,\ i=1,\dots,5$ are true. Set $\widehat{\mathbf{D}}=\widehat{\Sigma}_f^{-1}+\widehat{\mathbf{B}}(\widehat{\Sigma}_u^{\tau})^{-1}\widehat{\mathbf{B}}^{\mathsf{T}}$ and $\mathbf{D}=\Upsilon^{\mathsf{T}}\Sigma_f^{-1}\Upsilon+\Upsilon^{\mathsf{T}}\mathbf{B}(\Sigma_u^{\tau})^{-1}\mathbf{B}^{\mathsf{T}}\Upsilon$. By (2.7) and (2.8), we have

$$\begin{aligned}\widehat{\Sigma}_y^{-1}-\Sigma_y^{-1}&=\left((\widehat{\Sigma}_u^{\tau})^{-1}-\Sigma_u^{-1}\right)-\left((\widehat{\Sigma}_u^{\tau})^{-1}-\Sigma_u^{-1}\right)\widehat{\mathbf{B}}^{\mathsf{T}}\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{B}}(\widehat{\Sigma}_u^{\tau})^{-1}\\&\quad-\Sigma_u^{-1}\widehat{\mathbf{B}}^{\mathsf{T}}\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{B}}\left((\widehat{\Sigma_u}^{\tau})^{-1}-\Sigma_u^{-1}\right)-\Sigma_u^{-1}(\widehat{\mathbf{B}}^{\mathsf{T}}-\mathbf{B}^{\mathsf{T}}\Upsilon)\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{B}}\Sigma_u^{-1}\\&\quad-\Sigma_u^{-1}\mathbf{B}^{\mathsf{T}}\Upsilon\widehat{\mathbf{D}}^{-1}(\widehat{\mathbf{B}}-\Upsilon^{\mathsf{T}}\mathbf{B})\Sigma_u^{-1}+\Sigma_u^{-1}\mathbf{B}^{\mathsf{T}}\Upsilon\widehat{\mathbf{D}}^{-1}(\widehat{\mathbf{D}}-\mathbf{D})\mathbf{D}^{-1}\Upsilon^{\mathsf{T}}\mathbf{B}\Sigma_u^{-1}\\&=\Theta_1+\Theta_2+\Theta_3+\Theta_4+\Theta_5+\Theta_6,\end{aligned}$$

where the matrices $\Theta_i,\ i=1,\dots,6$ are defined in the obvious way. When the event V_5 happens, by Assumption A and Weyl's inequality, we have

$$c_0/2\le\lambda_{\min}(\Sigma_u)/2\le\lambda_{\min}(\widehat{\Sigma}_u)\le\lambda_{\max}(\widehat{\Sigma}_u)\le 2\lambda_{\max}(\Sigma_u)\le 2c_1. \tag{A.17}$$

Hence

$$\lambda_{\min}(\widehat{\mathbf{D}})\ge\lambda_{\min}(\widehat{\mathbf{B}\mathbf{B}}^{\mathsf{T}})\cdot\lambda_{\min}(\widehat{\Sigma}_u^{\tau})>pC$$

with $C=1/(2c_1)$. Similarly, $\lambda_{\min}(\mathbf{D})\ge pC$. We then have

$$\|\widehat{\mathbf{D}}^{-1}\|,\|\mathbf{D}^{-1}\|\le 1/pC. \tag{A.18}$$

When the events A_0, V_3 happen, via (A.17) and (A.18), we have

$$\|\Theta_1\|,\|\Theta_2\|,\|\Theta_3\|\le C\kappa(p,n). \tag{A.19}$$

By (A.11), (A.17) and (A.18), we have

$$\|\Theta_4\|,\|\Theta_5\|\le C\Big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{p}}\Big). \tag{A.20}$$

By the definition, we have

$$\begin{aligned}\widehat{\mathbf{D}}-\mathbf{D}&=(\widehat{\Sigma}_f^{-1}-\Sigma_f^{-1})+(\widehat{\mathbf{B}}-\Upsilon^{\mathsf{T}}\mathbf{B})(\widehat{\Sigma}_u^{\tau})^{-1}\widehat{\mathbf{B}}^{\mathsf{T}}\\&\quad+\Upsilon^{\mathsf{T}}\mathbf{B}(\widehat{\Sigma}_u^{\tau})^{-1}(\widehat{\mathbf{B}}^{\mathsf{T}}-\mathbf{B}^{\mathsf{T}}\Upsilon)+\Upsilon^{\mathsf{T}}\mathbf{B}((\widehat{\Sigma}_u^{\tau})^{-1}-\Sigma_u^{-1})\mathbf{B}^{\mathsf{T}}\Upsilon.\end{aligned}$$

Then by Assumption D, (A.11), (A.17) and when the events A_0, V_5 happen, we have

$$\begin{aligned}\|\widehat{\mathbf{D}}-\mathbf{D}\|&\le\|\widehat{\Sigma}_f^{-1}\|+\|\Sigma_f^{-1}\|+\|\widehat{\mathbf{B}}-\Upsilon^{\mathsf{T}}\mathbf{B}\|\cdot\|(\widehat{\Sigma}_u^{\tau})^{-1}\|\cdot\|\widehat{\mathbf{B}}\|\\&\quad+\|\Upsilon\|\cdot\|\mathbf{B}\|\cdot\|(\widehat{\Sigma}_u^{\tau})^{-1}\|\cdot\|\widehat{\mathbf{B}}^{\mathsf{T}}-\mathbf{B}^{\mathsf{T}}\Upsilon\|+\|\Upsilon\|^2\cdot\|\mathbf{B}\|^2\cdot\|(\widehat{\Sigma}_u^{\tau})^{-1}-\Sigma_u^{-1}\|\\&\le C+C(\sqrt{p}+\sqrt{n})+cp\kappa(p,n)\\&\le cp\kappa(p,n).\end{aligned} \tag{A.21}$$

It follows from (A.17), (A.18) and (A.21) that

$$\|\Theta_6\|\le\|\Sigma_u^{-1}\|^2\cdot\|\mathbf{B}\mathbf{B}^{\mathsf{T}}\|\cdot\|\Upsilon\|^2\cdot\|\widehat{\mathbf{D}}^{-1}\|^2\cdot\|\widehat{\mathbf{D}}-\mathbf{D}\|\le C\kappa(p,n). \tag{A.22}$$

Combining (A.19), (A.20) with (A.22), the proof of Theorem 2.2 is completed.

Proof of Theorem 2.3. Noting $\bar{\mathbf{y}}=\boldsymbol{\mu}+\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}}+\bar{\mathbf{u}}$, we have $\bar{\mathbf{y}}^{\mathsf{T}}\bar{\mathbf{y}}/p\le 3(\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}+\bar{\mathbf{f}}^{\mathsf{T}}\mathbf{B}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}}+\bar{\mathbf{u}}^{\mathsf{T}}\bar{\mathbf{u}})/p$. A simple calculation reveals that

$$E\Big(\frac{\bar{\mathbf{f}}^{\mathsf{T}}\mathbf{B}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}}}{p}\Big)=\frac{1}{n^2p}\sum_{t=1}^{n}E\Big(\mathrm{tr}\,\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{f}_t\mathbf{f}_t^{\mathsf{T}}\Big)\le\frac{1}{np}\mathrm{tr}\,\mathbf{B}\mathbf{B}^{\mathsf{T}}\Sigma_f\le\frac{C}{n},$$
$$E\Big(\frac{\bar{\mathbf{u}}^{\mathsf{T}}\bar{\mathbf{u}}}{p}\Big)=\frac{1}{n^2p}\sum_{t=1}^{n}E\Big(\mathrm{tr}\,\mathbf{u}_t\mathbf{u}_t^{\mathsf{T}}\Big)\le\frac{1}{np}\mathrm{tr}\,\Sigma_u\le\frac{C}{n}.$$

Hence, by Chebyshev's inequality, in probability

$$\frac{\bar{\mathbf{f}}^{\mathsf{T}}\mathbf{B}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}}}{p} \le \frac{m_n}{n}, \qquad \frac{\bar{\mathbf{u}}^{\mathsf{T}}\bar{\mathbf{u}}}{p} \le \frac{m_n}{n}, \tag{A.23}$$

where m_n is a sequence converging to infinity with $m_n/n \to 0$ as $n \to \infty$. Since $\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu}/p \le C$, we have $\bar{\mathbf{y}}^{\mathsf{T}}\bar{\mathbf{y}}/p \le C$ in probability. Thus, by Theorem 2.2,

$$\left\| \frac{\bar{\mathbf{y}}^{\mathsf{T}}\widehat{\Sigma}_y^{-1}\bar{\mathbf{y}}}{p} - \frac{\bar{\mathbf{y}}^{\mathsf{T}}\Sigma_y^{-1}\bar{\mathbf{y}}}{p} \right\| \le \frac{\bar{\mathbf{y}}^{\mathsf{T}}\bar{\mathbf{y}}}{p} \cdot \|\widehat{\Sigma}_y^{-1} - \Sigma_y^{-1}\| \le C\kappa(p, n), \tag{A.24}$$

in probability.

One the other hand, via (A.23), we have

$$\begin{aligned}
\frac{1}{p}\left|\bar{\mathbf{y}}^{\mathsf{T}}\Sigma_y^{-1}\bar{\mathbf{y}} - \boldsymbol{\mu}^{\mathsf{T}}\Sigma_y^{-1}\boldsymbol{\mu}\right| &= \frac{1}{p}\left|\bar{\mathbf{f}}^{\mathsf{T}}\mathbf{B}\Sigma_y^{-1}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}} + \bar{\mathbf{u}}^{\mathsf{T}}\Sigma_y^{-1}\bar{\mathbf{u}} + 2\boldsymbol{\mu}^{\mathsf{T}}\Sigma_y^{-1}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}} + 2\boldsymbol{\mu}^{\mathsf{T}}\Sigma_y^{-1}\bar{\mathbf{u}} + 2\bar{\mathbf{u}}^{\mathsf{T}}\Sigma_y^{-1}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}}\right| \\
&\le C\Big(\frac{1}{p}\bar{\mathbf{f}}^{\mathsf{T}}\mathbf{B}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}} + \frac{1}{p}\bar{\mathbf{u}}^{\mathsf{T}}\bar{\mathbf{u}} + \sqrt{\frac{1}{p}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu} \cdot \frac{1}{p}\bar{\mathbf{f}}^{\mathsf{T}}\mathbf{B}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}}} + \sqrt{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\mu} \cdot \frac{1}{p}\bar{\mathbf{u}}^{\mathsf{T}}\bar{\mathbf{u}}} + \sqrt{\frac{1}{p}\bar{\mathbf{f}}^{\mathsf{T}}\mathbf{B}\mathbf{B}^{\mathsf{T}}\bar{\mathbf{f}} \cdot \frac{1}{p}\bar{\mathbf{u}}^{\mathsf{T}}\bar{\mathbf{u}}}\Big) \\
&\le \sqrt{\frac{m_n}{n}}.
\end{aligned}$$

Therefore, if we let $m_n \le n\kappa^2(p, n)$, since $\sqrt{n}\kappa(p, n) \to \infty$, we then obtain in probability

$$\frac{1}{p}\left|\bar{\mathbf{y}}^{\mathsf{T}}\Sigma_y^{-1}\bar{\mathbf{y}} - \boldsymbol{\mu}^{\mathsf{T}}\Sigma_y^{-1}\boldsymbol{\mu}\right| \le C\kappa(p, n). \tag{A.25}$$

The proof of Theorem 2.3(i) is completed by (A.24) and (A.25). Similarly, we can prove Theorem 2.3(ii) and (iii).

References

[1] J. Bai, Inferential theory for factor models of large dimensions, Econometrica 71 (2003) 135–171.
[2] Z.D. Bai, H. Liu, W.K. Wong, Enhancement of the applicability of Markowitz's portfolio optimization by utilizing random matrix theory, Math. Finance 19 (2009) 639–667.
[3] J. Bai, S. Ng, Determining the number of factors in approximate factor models, Econometrica 70 (2002) 191–221.
[4] T. Cai, W. Liu, Adaptive thresholding for sparse covariance matrix estimation, J. Amer. Statist. Assoc. 106 (2011) 672–684.
[5] B.B. Chen, G.M. Pan, Convergence of the largest eigenvalue of normalized sample covariance matrices when p and n both tend to infinity with their ratio converging to zero, Bernoulli 18 (2012) 1405–1420.
[6] Y. Chen, A. Wiesel, A.O. Hero, Shrinkage estimation of high dimensional covariance matrices, in: Proc. of ICASSP-2009, 2009, pp. 2937–2940.
[7] J. Detemple, M. Rindisbacher, Closed-form solutions for optimal portfolio selection with stochastic interest rate and investment constraints, Math. Finance 15 (2005) 539–568.
[8] S. Emmer, C. Klüppelberg, R. Korn, Optimal portfolios with bounded capital at risk, Math. Finance 11 (2001) 365–384.
[9] J. Fan, Y. Fan, J. Lv, High dimensional covariance matrix estimation using a factor model, J. Econometrics 147 (2008) 186–197.
[10] J. Fan, Y. Liao, M. Mincheva, High-dimensional covariance matrix estimation in approximate factor models, Ann. Statist. 39 (2011) 3320–3356.
[11] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (2013) 603–680.
[12] G.M. Frankfurter, H.E. Phillips, J.P. Seagle, Performance of the Sharpe portfolio selection model: a comparison, J. Financ. Quant. Anal. 6 (1976) 195–204.
[13] M. Hallin, R. Liška, Determining the number of factors in the general dynamic factor model, J. Amer. Statist. Assoc. 102 (2007) 603–617.
[14] J.D. Jobson, B. Korkie, Estimation for Markowitz efficient portfolios, J. Amer. Statist. Assoc. 75 (1980) 544–554.
[15] X. Ju, N.D. Pearson, Using value-at-risk to control risk taking: how wrong can you be? J. Risk 1 (1999) 5–36.
[16] Y. Kroll, H. Levy, H.M. Markowitz, Mean–variance versus direct utility maximization, J. Finance 39 (1984) 47–61.
[17] L. Laloux, P. Cizeau, J.P. Bouchaud, M. Potters, Noise dressing of finacial corelation matrices, Phys. Rev. Lett. 83 (1999) 1467–1470.
[18] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2004) 365–411.
[19] R.A. Maller, D.A. Turkington, New light on the portfolio allocation problem, Math. Methods Oper. Res. 56 (2002) 501–511.
[20] H.M. Markowitz, Portfolio selection, J. Finance 7 (1952) 77–91.
[21] H.M. Markowitz, Portfolio Selection, John Wiley and Sons, New York, 1959.
[22] H.M. Markowitz, Portfolio Selection: Efficient Diversification of Investment, Blackwell, Cambridge, MA, 1991.
[23] J.R. McNamara, Portfolio selection using stochastic dominance criteria, Decis. Sci. 29 (1998) 785–801.
[24] R.C. Merton, An analytic derivation of the efficient portfolio frontier, J. Finan. Quant. Anal. 7 (1972) 1851–1872.
[25] R.O. Michaud, The Markowitz optimization enigma: is "optimized" optimal? Financ. Anal. J. 45 (1989) 31–42.
[26] K. Muthuraman, S. Kumar, Multidimensional portfolio optimization with proportional transaction costs, Math. Finance 16 (2006) 301–335.
[27] A. Onatski, Determining the number of factors from the empirical distribution of eigenvalues, Rev. Econ. Stat. 92 (2010) 1004–1016.
[28] G.M. Pan, Comparison between two types of large sample covariance matrices, Ann. Inst. H. Poincaré Probab. Statist. 50 (2014) 655–677.
[29] V.V. Petrov, Sums of Independent Random Variables, Springer, New York, 1975.
[30] S.A. Ross, The arbitrage theory of capital asset pricing, J. Econom. Theory 13 (1976) 341–360.
[31] J. Xia, Mean–variance portfolio choice: quadratic pratial hedging, Math. Finance 15 (2005) 533–538.
[32] J. Xia, J.A. Yan, Markowitz portfolios optimization in an incomplete market, Math. Finance 16 (2006) 203–216.