Journal of Multivariate Analysis 132 (2014) 215–228

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix

Taras Bodnar [a], Arjun K. Gupta [b,*], Nestor Parolya [c]

[a] Department of Mathematics, Humboldt-University of Berlin, D-10099 Berlin, Germany

[b] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA

[c] Institute of Empirical Finance (Econometrics), Leibniz University of Hannover, 30167, Germany

ARTICLE INFO

Article history:
Received 9 March 2013
Available online 4 September 2014

AMS 2010 subject classifications:
60B20
62H12
62G20
62G30

Keywords:
Large-dimensional asymptotics
Random matrix theory
Covariance matrix estimation

ABSTRACT

In this work we construct an optimal linear shrinkage estimator for the covariance matrix in high dimensions. The recent results from the random matrix theory allow us to find the asymptotic deterministic equivalents of the optimal shrinkage intensities and estimate them consistently. The developed distribution-free estimators obey *almost surely* the smallest Frobenius loss over all linear shrinkage estimators for the covariance matrix. The case we consider includes the number of variables $p \to \infty$ and the sample size $n \to \infty$ so that $p/n \to c \in (0, +\infty)$. Additionally, we prove that the Frobenius norm of the sample covariance matrix tends almost surely to a deterministic quantity which can be consistently estimated.

Published by Elsevier Inc.

1. Introduction

Nowadays, the estimation of the covariance matrix is one of the most important problems not only in statistics but also in finance, wireless communications, biology, etc. The traditional estimator of the covariance matrix, i.e. its sample counterpart, seems to be a good decision only when the dimension p is much smaller than the sample size n. This case is called the "standard asymptotics" (see, e.g., [21]). Here, the sample covariance matrix is proven to be an unbiased and a consistent estimator for the covariance matrix. More problems arise when p is comparable to n, i.e. both the dimension p and the sample size n tend to infinity while their ratio p/n tends to a positive constant c. It is called the "large dimensional asymptotics" or "Kolmogorov asymptotics" (see, e.g., [6,8]). This type of asymptotics have been exhaustively studied by Girko [15,16], where it was called the "general statistical analysis". There is a great amount of research done on the asymptotic behavior of functionals of the sample covariance matrix under the large dimensional asymptotics (see, e.g., [17–19,5]).

There are some significant improvements in the case when the covariance matrix has a special structure, e.g. sparse, low rank, etc. (see, [7,24,9,10], etc.). The case when the underlying random process obeys the factor structure is studied by Fan et al. [11]. In these cases the covariance matrix can be consistently estimated even in high-dimensional case. In the case when no additional information on the structure of the covariance matrix is available, the problem has not been studied in detail up to now. The exception is the paper of Ledoit and Wolf [22], where a linear shrinkage estimator was suggested which possesses the smallest Frobenius loss in quadratic mean.

* Corresponding author.
E-mail address: gupta@bgsu.edu (A.K. Gupta).

http://dx.doi.org/10.1016/j.jmva.2014.08.006
0047-259X/Published by Elsevier Inc.

Marčenko and Pastur [23], Yin [31], Silverstein [26], Bai et al. [2], Bai and Silverstein [5] used the large dimensional asymptotics to study the asymptotic behavior of the eigenvalues of general random matrices. They discovered that appropriately transformed random matrix at infinity has a nonrandom behavior and showed how to find the limiting density of its eigenvalues. In particular, Silverstein [26] proved under very general conditions that the Stieltjes transform of the sample covariance matrix tends almost surely to a nonrandom function which satisfies some equation. This equation was first derived by Marčenko and Pastur [23], who showed how the real covariance matrix and its sample estimate are connected at infinity. In our work we use this result for estimating functionals of the covariance matrix consistently.

In this work we concentrate on certain type of estimators, namely the shrinkage estimators. The shrinkage estimators were introduced by Stein [30]. They are constructed as a linear combination of the sample estimator and some known target. These estimators have remarkable property: they are biased but can significantly reduce the mean square error of the estimator. In the large as well as in the small dimensional cases it is difficult to find the consistent estimators for the so-called shrinkage intensities. In this situation Ledoit and Wolf [22] made progress when the target matrix is the identity and found a feasible linear shrinkage estimator for the covariance matrix which is optimal in the sense of the squared mean. This estimator provided a remarkable dominance over the sample estimator and other known estimators for the covariance matrix. The linear shrinkage presented by Ledoit and Wolf [22] shows its best performance in case when the eigenvalues of the covariance matrix are not dispersed and/or the concentration ratio c is large.

In this paper we extend the work of Ledoit and Wolf [22] by constructing a more general linear shrinkage estimator for a large dimensional covariance matrix. The target matrix here is considered to be an arbitrary symmetric positive definite matrix with uniformly bounded trace norm. Using random matrix theory we prove that the optimal shrinkage intensities are nonrandom at infinity, find their asymptotic deterministic equivalents and estimate them consistently. Additionally we show that the Frobenius norm of the covariance matrix tends to a deterministic quantity which can also be estimated consistently. The resulting estimator obeys *almost surely* the smallest Frobenius loss when the dimension p and the sample size n increase together and $p/n \to c \in (0, \infty)$ as $n \to \infty$.

The rest of paper is organized as follows. In Section 2 we present the preliminary results from the random matrix theory which are used in the proofs of the theorems. Section 3 contains the *oracle* linear shrinkage estimator and the main asymptotic results on the shrinkage intensities and the Frobenius norm of the sample covariance matrix. In Section 4 we present the *bona fide* linear shrinkage estimator for the covariance matrix and make a short comparison with the well-known Ledoit and Wolf [22] estimator. The results of the empirical study are provided in Section 5, while Section 6 summarizes all main results of the paper. The proofs of the theorems are moved to the Appendix.

2. Preliminary results and large dimensional asymptotics

By "large dimensional asymptotics" or "Kolmogorov asymptotics" it is understood that $\frac{p}{n} \to c \in (0, +\infty)$ where the number of variables $p \equiv p(n)$ and the sample size n both tend to infinity. In this case the traditional sample estimators perform poorly or very poorly and tend to over/underestimate the population covariance matrix.

We use the following notations in the paper:

- $\mathbf{\Sigma}_n$ stands for the covariance matrix, and $\mathbf{S}_n$ denotes the corresponding sample covariance matrix.
- The pairs $(\tau_i, \boldsymbol{v}_i)$ for $i = 1, \ldots, p$ are the collection of eigenvalues and the corresponding orthonormal eigenvectors of the covariance matrix $\mathbf{\Sigma}_n$.
- $H_n(t)$ is the empirical distribution function (e.d.f.) of the eigenvalues of $\mathbf{\Sigma}_n$, i.e.,

$$H_n(t) = \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}_{\{\tau_i < t\}} \tag{2.1}$$

  where $\mathbb{1}_{\{\cdot\}}$ is the indicator function.
- Let $\mathbf{X}_n$ be a $p \times n$ matrix which consists of independent and identically distributed (i.i.d.) real random variables with zero mean and unit variance such that

$$\mathbf{Y}_n = \mathbf{\Sigma}_n^{\frac{1}{2}} \mathbf{X}_n. \tag{2.2}$$

In the derivation of the main results the following five assumptions are used.

(A1) The population covariance matrix $\mathbf{\Sigma}_n$ is a nonrandom p-dimensional positive definite matrix.
(A2) Only the matrix $\mathbf{Y}_n$ is observable. We know neither $\mathbf{X}_n$ nor $\mathbf{\Sigma}_n$ itself.
(A3) We assume that $H_n(t)$ converges to some limit $H(t)$ at all points of continuity of H.
(A4) The elements of the matrix $\mathbf{X}_n$ have uniformly bounded moments of order $4 + \varepsilon$, $\varepsilon > 0$.
(A5) The largest eigenvalue of the covariance matrix $\mathbf{\Sigma}_n$ is at most of the order $O(\sqrt{p})$. Moreover, we assume that the order of only finite number of eigenvalues could depend on p.

The assumptions (A1)–(A3) are important to prove Marčenko–Pastur equation (see, e.g., [26]) and they are standard in the large dimensional asymptotics (see, e.g., [5]). In particular, the assumption (A3) on the existence of the limiting population

spectral function $H(t)$ is also very general because the support of $H(t)$ may be unbounded and non-compact. The assumption (A4) is needed for the proofs. This assumption is much weaker than the corresponding one used by Ledoit and Wolf [22] for similar purposes which is the existence of the 8th moment. The assumption (A5) ensures that the covariance matrix may have a couple of very large eigenvalues. Indeed, it is practically relevant and a good example of a model with unbounded eigenvalues of the covariance matrix is the factor model (see, [1,3,12]). Moreover, even if the structure of the data is more complex and the factor model is not present we still fall within the proposed theoretical assumptions (A1)–(A5). To the end, we do not impose the assumption of any particular distribution on the underlying data generating process and the presented framework is also applicable if the sample covariance matrix is singular, i.e. if the sample size n is larger than the dimension p. As a result, the assumptions (A1)–(A5) are general enough to cover many practical situations.

Let $(\lambda_i, \mathbf{u}_i)$ for $i = 1, \ldots, p$ denote the set of eigenvalues and the corresponding orthonormal eigenvectors of the sample covariance matrix[1]

$$\mathbf{S}_n = \frac{1}{n}\mathbf{Y}_n\mathbf{Y}_n' = \frac{1}{n}\mathbf{\Sigma}_n^{\frac{1}{2}}\mathbf{X}_n\mathbf{X}_n'\mathbf{\Sigma}_n^{\frac{1}{2}}. \tag{2.3}$$

Similar to (2.1) the (e.d.f.) of the sample covariance matrix $\mathbf{S}_n$ is defined by

$$F_n(\lambda) = \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{\lambda_i<\lambda\}}, \quad \lambda \in \mathbb{R}. \tag{2.4}$$

The most powerful tool for investigating the asymptotics of (e.d.f.) $F_n(\lambda)$ is the Stieltjes transform. For a nondecreasing function G with bounded variation the Stieltjes transform is defined by

$$\forall z \in \mathbb{C}^+ \quad m_G(z) = \int_{-\infty}^{+\infty} \frac{1}{\lambda - z} dG(\lambda). \tag{2.5}$$

In our notations $\mathbb{C}^+ = \{z \in \mathbb{C} : \mathbf{Im}(z) > 0\}$ is a half-plane of complex numbers with strictly positive imaginary part and any complex number is given by $z = \mathbf{Re}(z) + i\,\mathbf{Im}(z)$ with $\mathbf{Re}(z)$ and $\mathbf{Im}(z)$ the real and the imaginary parts accordingly. The Stieltjes transform and its importance for the behavior of spectrum of large dimensional random matrices is discussed in [27] in detail.

For all $z \in \mathbb{C}^+$ the Stieltjes transform of the sample (e.d.f.) $F_n(\lambda)$ is given by

$$m_{F_n}(z) = \frac{1}{p}\sum_{i=1}^{p}\int_{-\infty}^{+\infty} \frac{1}{\lambda - z}\delta(\lambda - \lambda_i)d\lambda = \frac{1}{p}\mathrm{tr}\{(\mathbf{S}_n - z\mathbf{I})^{-1}\}, \tag{2.6}$$

where $\mathbf{I}$ is a suitable identity matrix, $\mathrm{tr}(\cdot)$ is the trace of the matrix, and $\delta(\cdot)$ is the Dirac delta function.

Marčenko and Pastur [23] proved that the (e.d.f.) $F_n(\lambda)$ converges almost surely (a.s.) to a nonrandom limit $F(\lambda)$. Moreover, they derived an equation, the so-called Marčenko–Pastur (MP) equation, which shows the connection between $F(\lambda)$ and $H(\tau)$ at infinity. Under more general conditions the MP equation was considered by several authors (see, [31,29,28]). The most general case was presented by Silverstein [26] where the strong convergence under very general conditions was established. For illustration purposes, we summarize this result in Theorem 2.1.

Theorem 2.1 (*Silverstein [26]*)**.** *Assume that assumptions* (A1)–(A3) *are satisfied on the common probability space and* $\frac{p}{n} \to c \in (0, +\infty)$ *as* $n \to \infty$. *Then* $F_n(t) \overset{\text{a.s.}}{\Rightarrow} F(t)$ *as* $n \to \infty$. *Moreover, the Stieltjes transform of* F *satisfies the following equation*

$$m_F(z) = \int_{-\infty}^{+\infty} \frac{1}{\tau(1 - c - czm_F(z)) - z} dH(\tau), \tag{2.7}$$

in the sense that $m_F(z)$ *is the unique solution of* (2.7) *for all* $z \in \mathbb{C}^+$.

The MP equation (2.7) has a closed-form solution in several restricted cases. The famous Marčenko–Pastur law appears only when the covariance matrix $\mathbf{\Sigma}_n$ is the multiple of the identity matrix, i.e., $\mathbf{\Sigma}_n = \sigma\mathbf{I}$.

3. Optimal shrinkage estimator for covariance matrix

In this section we construct an optimal shrinkage estimator for the covariance matrix under high-dimensional asymptotics. This estimator is only *oracle* one, i.e., it depends on unknown quantities. The corresponding *bona fide* estimator is given in Section 4.

[1] It must be noted that the sample mean vector was omitted because the 1-rank matrix $\bar{\mathbf{x}}\bar{\mathbf{x}}'$ does not effect the asymptotic behavior of the spectrum of the sample covariance matrix (see, [5, Theorem A.44]).

Note that the linear shrinkage estimator of Ledoit and Wolf [22] has the smallest Frobenius loss only in quadratic mean. Consequently, it would be a great advantage to construct an estimator with almost sure smallest Frobenius loss. The general linear shrinkage estimator (GLSE) for the covariance matrix is given by

$$\widehat{\mathbf{\Sigma}}_{\mathrm{GLSE}} = \alpha_n \mathbf{S}_n + \beta_n \mathbf{\Sigma}_0 \quad \text{with } \|\mathbf{\Sigma}_0\|_{\mathrm{tr}} \leq M \tag{3.1}$$

where the symmetric positive definite matrix $\mathbf{\Sigma}_0$ has bounded trace norm at infinity, i.e., there exists $M > 0$ such that $\sup_n \|\mathbf{\Sigma}_0\|_{\mathrm{tr}} = \sup_n \mathrm{tr}(\mathbf{\Sigma}_0) \leq M$. No assumption is imposed on the shrinkage intensities α_n and β_n which are objects of our interest.

Since the shrinkage intensities are the main object of investigation and $\mathbf{\Sigma}_0$ is fixed, a question arises how to choose this target matrix. It could depend on the underlying data as well as on the branch of science where the shrinkage estimator for the covariance matrix is applied, i.e., wireless communications, finance, etc. On the other hand, from the view of Bayesian statistics $\mathbf{\Sigma}_0$ can be interpreted as a hyperparameter of *a priori* distribution. The useful comments about the choice of the target matrix $\mathbf{\Sigma}_0$ can be found e.g. in [4,22].

The aim is to find the optimal shrinkage intensities α_n and β_n which minimize the Frobenius norm for a given nonrandom target matrix $\mathbf{\Sigma}_0$ expressed as

$$L_F^2 = \|\widehat{\mathbf{\Sigma}}_{\mathrm{GLSE}} - \mathbf{\Sigma}_n\|_F^2 = \|\mathbf{\Sigma}_n\|_F^2 + \|\widehat{\mathbf{\Sigma}}_{\mathrm{GLSE}}\|_F^2 - 2\,\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{\mathrm{GLSE}}\mathbf{\Sigma}_n). \tag{3.2}$$

As a result, using (3.1) we want to solve the following optimization problem

$$\begin{aligned} L_F^2 &= \alpha_n^2 \|\mathbf{S}_n\|_F^2 + 2\alpha_n\beta_n\,\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0) + \beta_n^2\|\mathbf{\Sigma}_0\|_F^2 \\ &\quad - 2\alpha_n\,\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_n) - 2\beta_n\,\mathrm{tr}(\mathbf{\Sigma}_n\mathbf{\Sigma}_0) \longrightarrow \min \quad \text{with respect to } \alpha_n \text{ and } \beta_n. \end{aligned} \tag{3.3}$$

Taking the derivatives of L_F^2 with respect to α_n and β_n and setting them equal to zero we get

$$\frac{\partial L_F^2}{\partial \alpha_n} = \alpha_n\|\mathbf{S}_n\|_F^2 + \beta_n\,\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0) - \mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_n) = 0, \tag{3.4}$$

$$\frac{\partial L_F^2}{\partial \beta_n} = \alpha_n\,\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0) + \beta_n\|\mathbf{\Sigma}_0\|_F^2 - \mathrm{tr}(\mathbf{\Sigma}_n\mathbf{\Sigma}_0) = 0. \tag{3.5}$$

The Hessian of the L_F^2 has the form

$$\mathbf{H} = \begin{pmatrix} \|\mathbf{S}_n\|_F^2 & \mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0) \\ \mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0) & \|\mathbf{\Sigma}_0\|_F^2 \end{pmatrix}. \tag{3.6}$$

From the following inequality it follows that the Hessian matrix $\mathbf{H}$ is always positive definite:

$$\begin{aligned} \det(\mathbf{H}) &= \|\mathbf{S}_n\|_F^2\|\mathbf{\Sigma}_0\|_F^2 - (\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0))^2 \geq \|\mathbf{S}_n\|_F^2\|\mathbf{\Sigma}_0\|_F^2 - \|\mathbf{S}_n\|_\infty^2(\mathrm{tr}(\mathbf{\Sigma}_0))^2 \\ &\overset{\text{Jensen}}{\geq} (\|\mathbf{S}_n\|_F^2 - \|\mathbf{S}_n\|_\infty^2)\|\mathbf{\Sigma}_0\|_F^2 > 0, \end{aligned} \tag{3.7}$$

where $\|\mathbf{S}_n\|_\infty$ denotes the spectral norm (square root of maximum eigenvalue of matrix $\mathbf{S}_n^2$). The last inequality in (3.7) is well-known (see, e.g., [20]).

From Eqs. (3.4) and (3.5) it is easy to find the optimal shrinkage intensities α_n^* and β_n^* as

$$\alpha_n^* = \frac{\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_n)\|\mathbf{\Sigma}_0\|_F^2 - \mathrm{tr}(\mathbf{\Sigma}_n\mathbf{\Sigma}_0)\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0)}{\|\mathbf{S}_n\|_F^2\|\mathbf{\Sigma}_0\|_F^2 - \left(\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0)\right)^2}, \tag{3.8}$$

$$\beta_n^* = \frac{\mathrm{tr}(\mathbf{\Sigma}_n\mathbf{\Sigma}_0)\|\mathbf{S}_n\|_F^2 - \mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_n)\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0)}{\|\mathbf{S}_n\|_F^2\|\mathbf{\Sigma}_0\|_F^2 - \left(\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0)\right)^2}. \tag{3.9}$$

Next, we consider the asymptotic behavior of the quantities (3.8) and (3.9), namely we look for their asymptotic equivalents. Recall that the sequence of random variables $\tilde{\xi}_n$ is called asymptotically equivalent to a nonrandom sequence ξ_n when

$$\left|\tilde{\xi}_n - \xi_n\right| \longrightarrow 0 \quad \text{a.s. for } n \to \infty. \tag{3.10}$$

Note that it is sufficient to know the asymptotic equivalents of the quantities $\|\mathbf{S}_n\|_F^2$, $\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_n)$ and $\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0)$. It is not difficult to find the asymptotic equivalents to the last two quantities. This is done in Theorem 3.2. More difficult is to find a asymptotic equivalent to the first quantity, namely $\|\mathbf{S}_n\|_F^2$. Since the Frobenius norm of the sample covariance matrix is very important in high dimensional statistics, it would be a great advantage to investigate its asymptotic behavior.

In Theorem 3.1 we present our first result where we show that the normalized Frobenius norm of $\mathbf{S}_n$ tends almost surely to a nonrandom quantity for $p/n \to c \in (0, +\infty)$ as $n \to \infty$.

Theorem 3.1. *Assume that* (A1)–(A5) *hold and* $\frac{p}{n} \to c \in (0, +\infty)$ *for* $n \to \infty$. *Then the Frobenius norm of the sample covariance matrix* $\phi_n = \frac{1}{p}\|\mathbf{S}_n\|_F^2$ *almost surely tends to a nonrandom variable* ϕ *which is given by*

$$\phi = \int_{-\infty}^{+\infty} \tau^2 dH(\tau) + c\left(\int_{-\infty}^{+\infty} \tau dH(\tau)\right)^2, \tag{3.11}$$

where $H(t)$ *denotes the limiting function of the spectral e.d.f.* $H_n(t)$ *defined in* (2.1).

The proof of Theorem 3.1 follows from the MP equation (2.7) and is presented in the Appendix. In particular, it ensures that the Frobenius norm of the sample covariance matrix is fixed and depends solely on the function H and c on infinity. Nevertheless, we are not able to estimate $\|\mathbf{\Sigma}_n\|_F^2$ using this knowledge since the function $H(t)$ is unknown. That is why we need to find the asymptotically equivalent quantity to $\|\mathbf{S}_n\|_F^2$.

Theorem 3.1 gives us some intuition about this quantity. The idea is to replace the integrals in (3.11) by the corresponding finite sums, namely $1/p\sum_{i=1}^p \tau_i^2 = 1/p\,\mathrm{tr}(\mathbf{\Sigma}_n^2)$ and $1/p\sum_{i=1}^p \tau_i = 1/p\,\mathrm{tr}(\mathbf{\Sigma}_n)$. The main advantage of the procedure is that this substitution does not effect the almost sure convergence.

Theorem 3.2. *Under assumptions* (A1)–(A5) *for* $\frac{p}{n} \to c \in (0, +\infty)$ *it holds that*

$$\frac{1}{p}\left|\|\mathbf{S}_n\|_F^2 - \left(\|\mathbf{\Sigma}_n\|_F^2 + \frac{c}{p}\|\mathbf{\Sigma}_n\|_{\mathrm{tr}}^2\right)\right| \longrightarrow 0 \quad \textit{a.s. for } n \to \infty \tag{3.12}$$

where $\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2 = \left(\mathrm{tr}(\mathbf{\Sigma}_n)\right)^2$ *is the squared trace norm of matrix* $\mathbf{\Sigma}_n$.

Additionally, for the quantity $\mathrm{tr}(\mathbf{S}_n\mathbf{\Theta})$, *where* $\mathbf{\Theta}$ *is a symmetric positive definite matrix with bounded trace norm, it holds that*

$$\frac{1}{p}\left|\mathrm{tr}(\mathbf{S}_n\mathbf{\Theta}) - \mathrm{tr}(\mathbf{\Sigma}_n\mathbf{\Theta})\right| \longrightarrow 0 \quad \textit{a.s. for } \frac{p}{n} \to c \in (0, +\infty) \textit{ as } n \to \infty. \tag{3.13}$$

The proof of the theorem is given in the Appendix. In contrast to Theorem 3.1, Theorem 3.2 contains a better interpretation of the asymptotic result (3.12). It shows, in particular, that the consistent estimator of the Frobenius norm of the real covariance matrix, $\|\mathbf{\Sigma}_n\|_F^2$, is not equal to its sample counterpart. On the other hand the functionals of the type $\mathrm{tr}(\mathbf{S}_n\mathbf{\Theta})$ are consistently estimated by the sample counterparts.

Moreover, it appears that under the large dimensional asymptotics, the Frobenius norm of $\mathbf{S}_n$ is shifted by the constant $c/p\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2$. As a result, we know the asymptotic value of the bias and the consistent estimator for Frobenius norm of the covariance matrix can be constructed by subtracting it from the sample counterpart. The exact form of this estimator is presented in Section 4.

Next, we consider the shrinkage intensities α_n^* and β_n^*. The application of Theorem 3.2 allows us to find their asymptotic properties. This is done in Corollary 3.1.

Corollary 3.1. *Assume that* (A1)–(A5) *are fulfilled. Then for* $\frac{p}{n} \to c \in (0, +\infty)$ *as* $n \to \infty$ *the optimal shrinkage intensities* α_n^* *and* β_n^* *satisfy*

$$\left|\alpha_n^* - \alpha^*\right| \longrightarrow 0 \quad \textit{a.s. for } n \to \infty, \tag{3.14}$$

where

$$\alpha^* = 1 - \frac{\frac{c}{p}\|\mathbf{\Sigma}\|_{\mathrm{tr}}^2\|\mathbf{\Sigma}_0\|_F^2}{\left(\|\mathbf{\Sigma}_n\|_F^2 + \frac{c}{p}\|\mathbf{\Sigma}_n\|_{\mathrm{tr}}^2\right)\|\mathbf{\Sigma}_0\|_F^2 - \left(\mathrm{tr}(\mathbf{\Sigma}_n\mathbf{\Sigma}_0)\right)^2} \tag{3.15}$$

and

$$\left|\beta_n^* - \beta^*\right| \longrightarrow 0 \quad \textit{a.s. for } n \to \infty, \tag{3.16}$$

where

$$\beta^* = \frac{\mathrm{tr}(\mathbf{\Sigma}_n\mathbf{\Sigma}_0)}{\|\mathbf{\Sigma}_0\|_F^2}\left(1 - \alpha^*\right). \tag{3.17}$$

Proof. This result follows straightforwardly from Theorem 3.2. Only the boundedness of the trace norm of the matrix $1/p\mathbf{\Sigma}_n$ is needed. It obviously follows from (A5), namely

$$\|1/p\mathbf{\Sigma}_n\|_{\mathrm{tr}} = 1/p\,\mathrm{tr}(\mathbf{\Sigma}_n) < \infty. \quad \square \tag{3.18}$$

First, we observe that the application of Corollary 3.1 and Theorem 3.2 ensures that the shrinkage intensities α^* and β^* can be consistently estimated. We summarize this result in Section 4.

The second interpretation of the results of Corollary 3.1 shows that $\alpha^* < 1$ as soon as $c > 0$ which follows directly from (3.7). Only if $c = 0$, we get $\alpha^* = 1$ and $\beta^* = 0$. In this case, the sample covariance matrix is a good estimator for the population covariance matrix which minimizes the Frobenius norm. In contrast if p increases, i.e. if $c > 0$, the general linear shrinkage estimator (3.1) improves the performance of the sample estimator. Moreover, the impact of this improvement becomes larger as p approaches n.

4. Bona fide estimator for the covariance matrix

This section is dedicated to the *bona fide* estimation of the unknown parameters from Section 3. As we have already mentioned the asymptotic shrinkage intensities α^* and β^* from Corollary 3.1 can be consistently estimated using the result of Theorem 3.2.

Obviously, both α^* and β^* depend on the Frobenius norm of the covariance matrix $\mathbf{\Sigma}_n$ and on the functionals of the type $\mathrm{tr}(\mathbf{\Sigma}_n\mathbf{\Theta})$. Due to Theorem 3.2 the consistent estimator of latter term is its sample counterpart, while the consistent estimator of the Frobenius norm $\psi_n = \frac{1}{p}\|\mathbf{\Sigma}_n\|_F^2$ is given by

$$\widehat{\psi}_n = \frac{1}{p}\|\mathbf{S}_n\|_F^2 - \frac{1}{np}\|\mathbf{S}_n\|_{\mathrm{tr}}^2. \tag{4.1}$$

The resulting estimator (4.1) is similar to that obtained by Girko [16]. In contrast to Girko [16], Theorem 3.2 was proved under more general conditions and we use the complex Stieltjes transform while Girko [16] used the real one. Girko's result holds for $c \in (0, 1)$ while our for $c \in (0, +\infty)$. Nevertheless, using Theorems 3.1 and 3.2 it can be shown that the so-called G_4^2-estimator considered by Girko [16] coincides with (4.1) and it is indeed the consistent estimator for the Frobenius norm of the covariance matrix under the large dimensional asymptotics.

The optimal linear shrinkage estimator (OLSE) for the covariance matrix $\mathbf{\Sigma}_n$ is given by

$$\widehat{\mathbf{\Sigma}}_{\mathrm{OLSE}} = \hat{\alpha}^*\mathbf{S}_n + \hat{\beta}^*\mathbf{\Sigma}_0 \quad \text{with } \|\mathbf{\Sigma}_0\|_{\mathrm{tr}} \le M, \tag{4.2}$$

where

$$\hat{\alpha}^* = 1 - \frac{\frac{1}{n}\|\mathbf{S}_n\|_{\mathrm{tr}}^2\|\mathbf{\Sigma}_0\|_F^2}{\|\mathbf{S}_n\|_F^2\|\mathbf{\Sigma}_0\|_F^2 - \left(\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0)\right)^2} \tag{4.3}$$

and

$$\hat{\beta}^* = \frac{\mathrm{tr}(\mathbf{S}_n\mathbf{\Sigma}_0)}{\|\mathbf{\Sigma}_0\|_F^2}\left(1 - \hat{\alpha}^*\right). \tag{4.4}$$

The OLSE estimator possesses almost surely the smallest Frobenius loss according to Theorem 3.2 and has a simple structure. Moreover, when $p > n$ and the sample covariance matrix $\mathbf{S}_n$ is singular the optimal linear shrinkage $\widehat{\mathbf{\Sigma}}_{\mathrm{OLSE}}$ stays invertible and applicable in practice.

Next, we consider an interesting special case when $\mathbf{\Sigma}_0 = \frac{1}{p}\mathbf{I}$ in more detail. In this case $\widehat{\mathbf{\Sigma}}_{\mathrm{OLSE}}$ looks very similar to the linear shrinkage estimator proposed by Ledoit and Wolf [22]. However, they are not equal.

First, the linear shrinkage estimator of Ledoit and Wolf [22] has the smallest Frobenius loss in quadratic mean while the suggested estimator in (4.2) ensures almost sure convergence to the oracle. Moreover, Ledoit and Wolf [22] assumed the existence of 8th moment while our estimator is derived under the assumption that $4 + \varepsilon$, $\varepsilon > 0$, moment exists.

Second, the estimator of Ledoit and Wolf [22] and the suggested optimal linear shrinkage estimator (4.2) with $\mathbf{\Sigma}_0 = 1/p\mathbf{I}$ differ in $\hat{\alpha}^*$. Instead of $\frac{1}{n}\|\mathbf{S}_n\|_{\mathrm{tr}}^2$ Ledoit and Wolf [22] used $\frac{1}{n^2}\sum_{i=1}^n \|\mathbf{y}_i\mathbf{y}_i' - \mathbf{S}_n\|_F^2$, where $\mathbf{y}_i$ are the columns of the matrix $\mathbf{\Sigma}_n^{1/2}\mathbf{X}_n$. Indeed, let $d^2 = \frac{1}{p}\|\mathbf{S}_n\|_F^2 - \left(1/p\,\mathrm{tr}(\mathbf{S}_n)\right)^2$ and $b^2 = \frac{1}{p}\frac{1}{n^2}\sum_{i=1}^n \|\mathbf{y}_i\mathbf{y}_i' - \mathbf{S}_n\|_F^2$. Then the estimate for α^* by Ledoit and Wolf [22] is given by

$$\hat{\alpha}_{LW}^* = 1 - \frac{\min\{b^2, d^2\}}{d^2}. \tag{4.5}$$

Consequently, from (4.5) we observe that the Ledoit–Wolf (LW) linear shrinkage estimator is constrained whereas our optimal shrinkage estimator (4.2) for $\mathbf{\Sigma}_0 = \frac{1}{p}\mathbf{I}$ is unconstrained. Moreover, if $b^2 > d^2$ in (4.5) then $\alpha_{LW}^* = 0$ independently how large p is with respect to n. In this case LW estimator is equal to the target matrix $\mathrm{tr}(\mathbf{S}_n)\frac{1}{p}\mathbf{I}$. In contrast, for the suggested

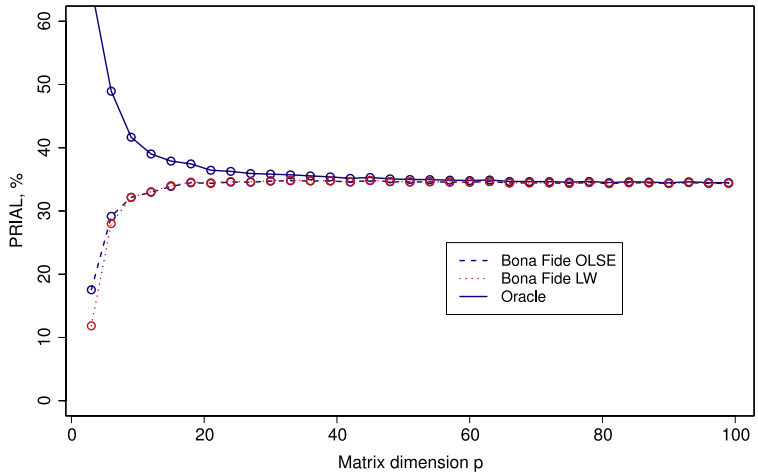

Fig. 1. PRIALs for the oracle, the bona fide OLSE, and the bona fide LW estimators for $p = 3k,\ k \in \{1, \ldots, 33\}$, $c = 1/3$, and $\mathbf{\Sigma}_0 = \frac{1}{p}\mathbf{I}$, 1000 repetitions.

OLSE estimator (4.2) it holds always that $0 < \alpha^* \leq 1$. Moreover, $\alpha^* = 1$ iff $c = 0$ which means that the sample covariance matrix possesses the smallest Frobenius loss only if p is much smaller than n. For $c > 0$, the sample covariance matrix is not an optimal estimator for the covariance matrix.

Third, the LW linear shrinkage estimator seems to be more computationally intensive than (4.2) for $\mathbf{\Sigma}_0 = \frac{1}{p}\mathbf{I}$. The reason is that the quantity b^2 is computed by a loop while $\frac{1}{n}\|\mathbf{S}_n\|_{\mathrm{tr}}^2$ needs only the computation of the trace.

Ledoit and Wolf [22] proved that their linear shrinkage estimator tends on average to the oracle one while our estimator tends to the oracle almost surely and, consequently, it is expected that they are asymptotically the same. Moreover, the choice $\mathbf{\Sigma}_0 = \frac{1}{p}\mathbf{I}$ can be too conservative and better results can be obtained for other values of $\mathbf{\Sigma}_0$.

In Fig. 1 we present the simulation results for normally distributed data in the case when $\mathbf{\Sigma}_0 = \frac{1}{p}\mathbf{I}$. We take, without loss of generality, $\mathbf{\Sigma}_n$ as a diagonal matrix and separate its spectrum in three equal parts with eigenvalues 0.1, 5 and 10. In terms of the corresponding cumulative distribution function of the eigenvalues of $\mathbf{\Sigma}_n$ (ct. Section 2) it holds that

$$H_n^{\mathbf{\Sigma}_n}(t) = 1/3\delta_{[0.1,\infty)}(t) + 1/3\delta_{[5,\infty)}(t) + 1/3\delta_{[10,\infty)}(t), \tag{4.6}$$

where δ is the Dirac delta function. Doing so we leave the structure of population covariance matrix unchanged for all dimensions p. Then, we compare the LW linear shrinkage estimator with the suggested OLSE estimator in terms of their PRIAL's. For an arbitrary estimator of the covariance matrix, $\widehat{\mathbf{M}}$, the PRIAL (Percentage Relative Improvement in Average Loss) is defined as

$$\mathrm{PRIAL}(\widehat{\mathbf{M}}) = \left(1 - \frac{E\|\widehat{\mathbf{M}} - \mathbf{\Sigma}_n\|_F^2}{E\|\mathbf{S}_n - \mathbf{\Sigma}_n\|_F^2}\right) \cdot 100\%. \tag{4.7}$$

By definition (4.7), $\mathrm{PRIAL}(\mathbf{S}_n)$ is equal to zero and $\mathrm{PRIAL}(\mathbf{\Sigma}_n)$ is equal to 100%.

Fig. 1 clearly shows that both the LW estimator and the suggested OLSE estimator converge quickly to their common oracle in average. Moreover we conclude that there is no significant difference between the two estimators.

In Fig. 2 we show how the prior knowledge of the structure of the population covariance matrix can improve the OLSE estimator when $\mathbf{\Sigma}_0 = 1/p\mathbf{I}$. We assume that the prior matrix $\mathbf{\Sigma}_0$ conforms to the spectrum separation of the covariance matrix $\mathbf{\Sigma}_n$. Thus, suppose we know that the spectrum of population covariance matrix is separated in three equal blocks (see equality (4.6)) and we do not take any other information into account. The diagonal elements of prior matrix $\mathbf{\Sigma}_0$ are chosen to be 1, 2 and 3. In terms of the cumulative distribution function of $\mathbf{\Sigma}_0$ it holds that

$$H^{\mathbf{\Sigma}_0}(t) = 1/3\delta_{[1,\infty)}(t) + 1/3\delta_{[2,\infty)}(t) + 1/3\delta_{[3,\infty)}(t). \tag{4.8}$$

In this case Fig. 2 shows that improvement of the OLSE estimator is about 40% over the OLSE estimator constructed with $\mathbf{\Sigma}_0 = 1/p\mathbf{I}$. It is remarkable that both the oracle and bona fide OLSE estimators dominate the corresponding oracle and bona fide estimators calculated for $\mathbf{\Sigma}_0 = 1/p\mathbf{I}$. Note that both the estimators now possess different oracles due to the prior $\mathbf{\Sigma}_0$.

Similar situation can be observed in the case $c = 2$. Here the proportion of zero eigenvalues is asymptotically equal to $1-c^{-1} = 0.5$, i.e. 50% (see, e.g. [5]). In Figs. 3 and 4 we present the corresponding results of the Monte Carlo simulations. Fig. 3 shows that the derived estimator for the population covariance matrix with the prior $\mathbf{\Sigma}_0 = 1/p\mathbf{I}$ is on average asymptotically the same as the linear shrinkage of Ledoit and Wolf [22]. It is noted that the overall performance of the OLSE estimator is significantly better than in the case $c = 1/3$.

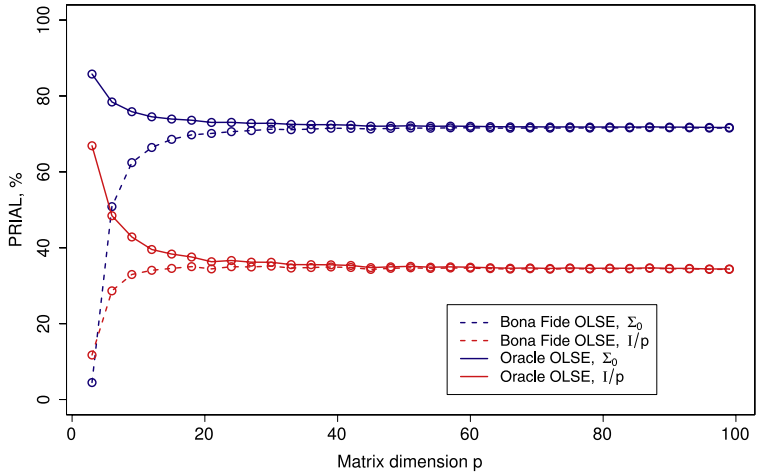

Fig. 2. PRIALs for the oracle and the bona fide OLSE estimators with $\mathbf{\Sigma}_0 = 1/p\mathbf{I}$ and $\mathbf{\Sigma}_0$ as given in (4.8) for $p = 3k,\ k \in \{1, \ldots, 33\}$ and $c = 1/3$, 1000 repetitions.

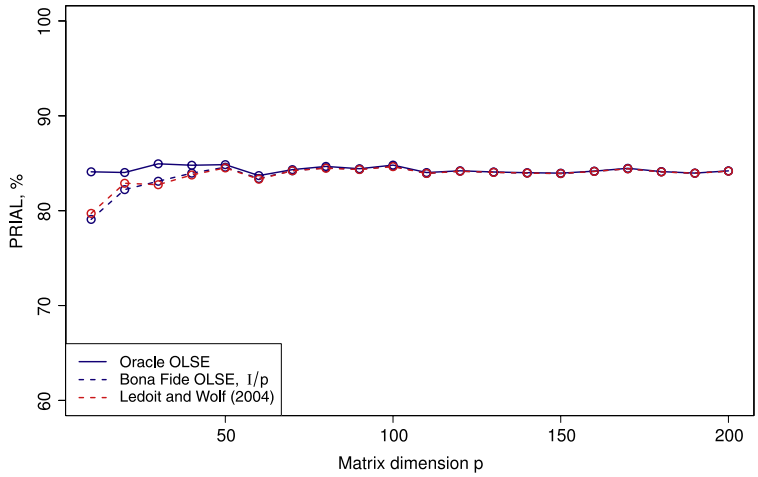

Fig. 3. PRIALs for the oracle, the bona fide OLSE, and the bona fide LW estimators for $p = 10k,\ k \in \{1, \ldots, 20\}$, $c = 2$, and $\mathbf{\Sigma}_0 = \frac{1}{p}\mathbf{I}$, 1000 repetitions.

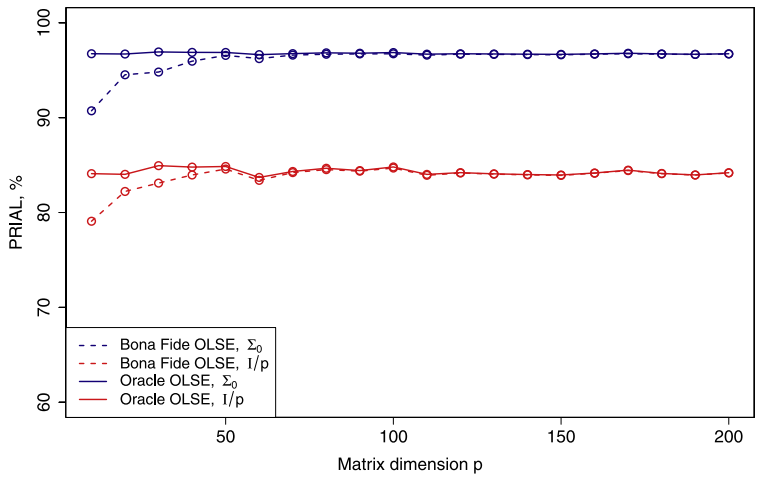

Fig. 4. PRIALs for the oracle and the bona fide OLSE estimators with $\mathbf{\Sigma}_0 = 1/p\mathbf{I}$ and $\mathbf{\Sigma}_0$ as given in (4.8) for $p = 10k,\ k \in \{1, \ldots, 20\}$ and $c = 2$, 1000 repetitions.

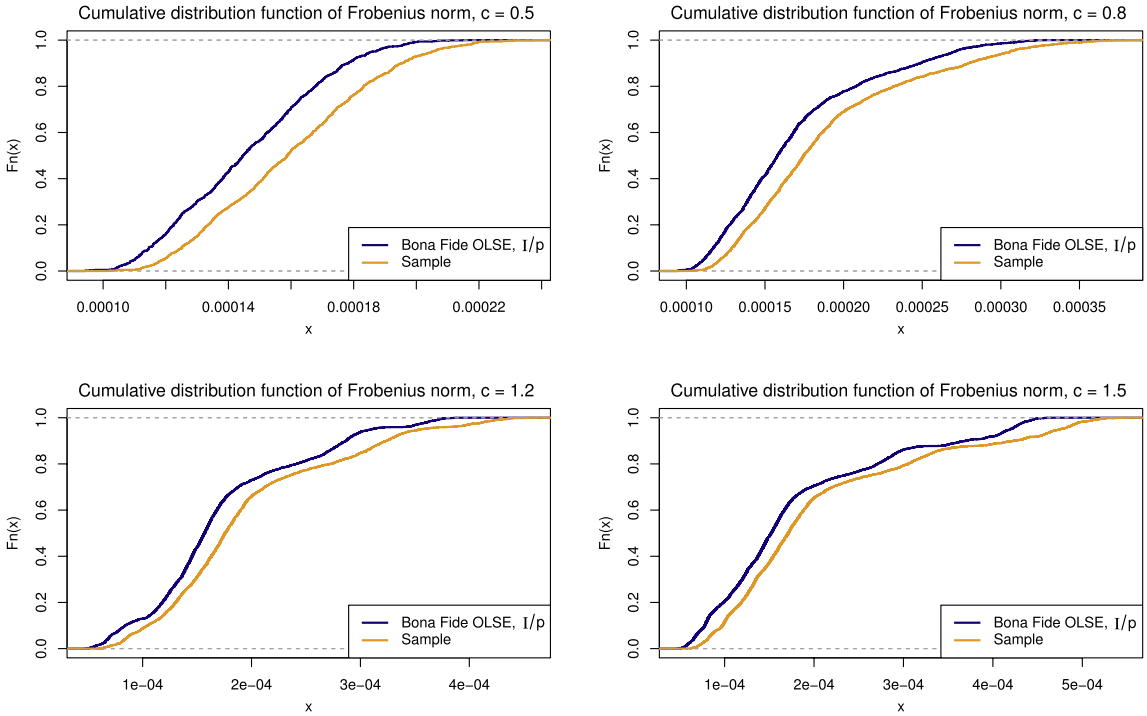

Fig. 5. Empirical distribution function of the Frobenius norms of the OLSE and the sample covariance matrix.

Fig. 4 presents the case when the additional information about the spectrum separation from (4.8) is taken into account. Here, the spectrum of $\mathbf{\Sigma}_0$ is determined by

$$H^{\mathbf{\Sigma}_0}(t) = 1/3\delta_{[1,\infty)}(t) + 1/3\delta_{[2,\infty)}(t) + 1/3\delta_{[60,\infty)}(t).$$

The dominance of about 20% is detected. It means that the information about the spectrum separation of the population covariance matrix plays a crucial role for choosing the prior target matrix for the OLSE estimator. Consequently, when the target matrix $\mathbf{\Sigma}_0$ departs away from the identity matrix, it has a positive influence on the OLSE estimator and can improve it significantly.

Finally, we note that the optimal linear shrinkage estimator (4.2) seems to be a good generalization and modification of the linear shrinkage estimator presented by Ledoit and Wolf [22], which is also a good alternative to the sample covariance matrix. This continues to be true even when the dimension p is greater than the sample size n.

5. Empirical study

In this section, we apply the derived estimator for the covariance matrix to real data which consist of daily asset returns on the 431 assets listed in the S&P 500 (Standard & Poor's 500) index and traded during the whole period from 13.01.2004 to 10.01.2014. The S&P 500 is based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ. The number of the considered assets reflects the common setting in high-dimensional portfolio problems.

Next we analyze the influence of the portfolio size and the sample size on the behavior of the estimators for the Frobenius norm and the maximum and minimum eigenvalues of the population covariance matrix which are based on the derived OLSE and the sample covariance matrix. In Fig. 5 we present the results in case of $p = 156$ and $n \in \{104, 130, 195, 312\}$ which leads to $c \in \{0.5, 0.8, 1.2, 1.5\}$, respectively. Since the choice of the assets is not unique, here we sample randomly $p = 156$ assets out of 431 and generate 10^3 different portfolios. In Figs. 6 and 7, the results are shown for several values of $p \in \{50, 100, 200, 300\}$. The sample size n is chosen such that $c \in (0, 3)$. For each point from the figure the maximum (minimum) eigenvalue is calculated based on a randomly chosen portfolio of the dimension p.

Fig. 5 provides the empirical distribution of the Frobenius norms of the OLSE and the sample covariance matrix in case of 10^3 randomly chosen portfolios of the dimension $p = 156$. Here, we observe that the Frobenius norm of the OLSE is uniformly smaller than the one of the sample covariance matrix. This result holds true for all of the considered values of $c \in \{0.5, 0.8, 1.2, 1.5\}$. It is in line with the results of Theorem 3.2 where it is proved that the Frobenius norm of the sample covariance matrix overestimates asymptotically the corresponding population value.

In Figs. 6 and 7, we analyze the behavior of the smallest and the largest eigenvalues of the OLSE and the sample estimator of the covariance matrix. The results in case of the smallest eigenvalue are of great importance in Finance since the smallest

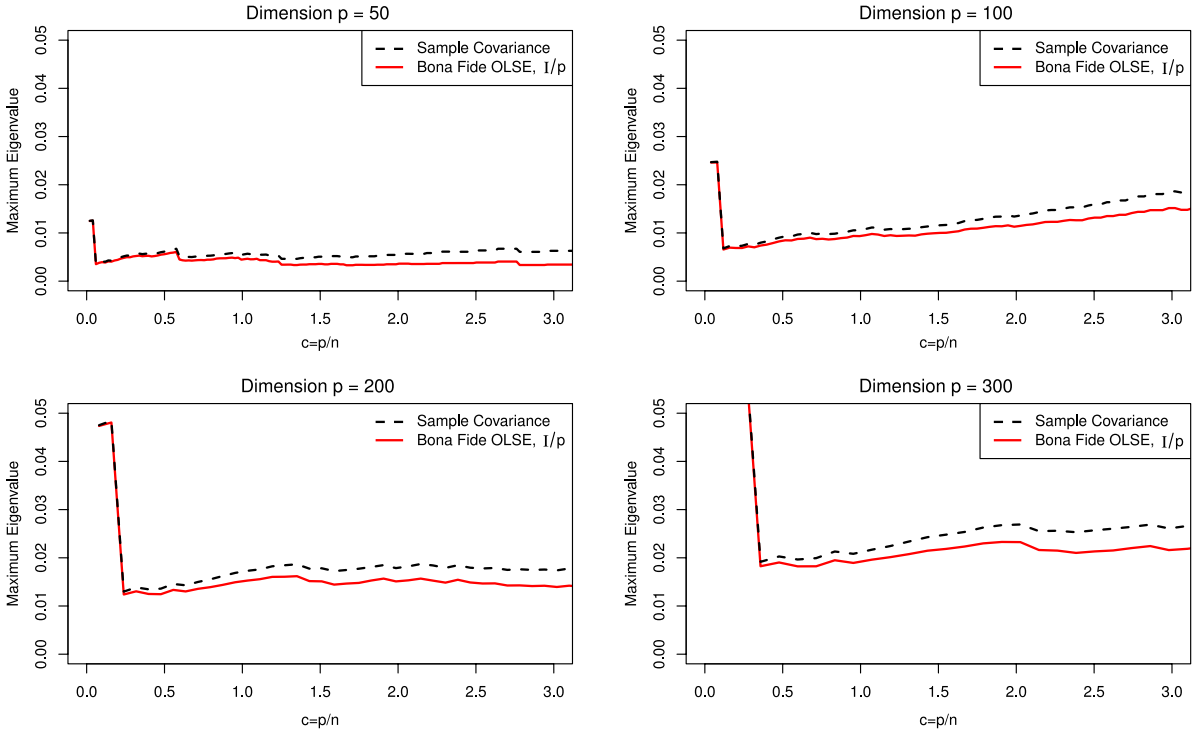

Fig. 6. Maximum eigenvalue of the OLSE and the sample covariance matrix.

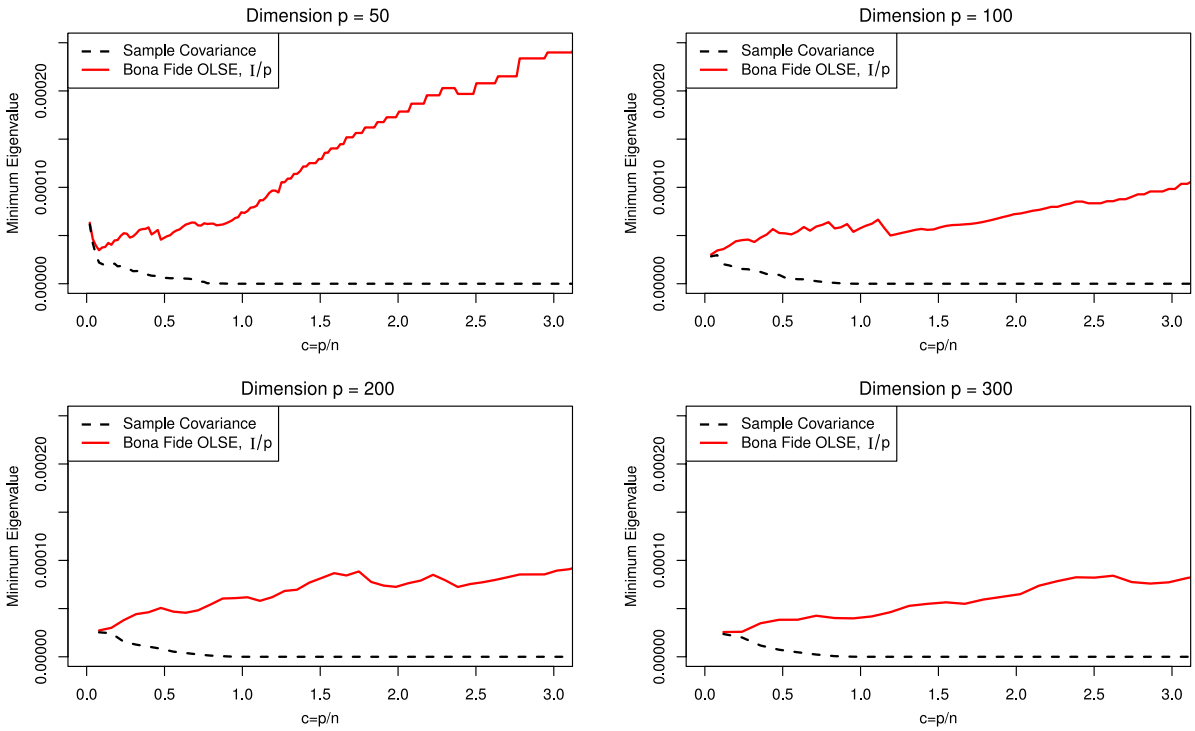

Fig. 7. Minimum eigenvalue of the OLSE and the sample covariance matrix.

eigenvalue of the covariance matrix is directly related to the variance of the global minimum variance portfolio which is very popular in Portfolio Theory (cf. [14]). We observe that the OLSE makes the largest eigenvalue smaller, whereas the smallest

eigenvalue becomes larger in comparison to the sample covariance matrix. This result allows us to correct the investor's overoptimism concerning the risk when the global minimum variance is constructed by using the sample covariance matrix.

6. Summary

This paper considers the problem of estimation of the covariance matrix in large dimensions. This case is known in the literature as the large dimensional asymptotics and it includes the number of variables $p \to \infty$ and the sample size $n \to \infty$ so that $p/n \to c \in (0, +\infty)$. Here, we construct the optimal linear shrinkage estimator (OLSE) for the covariance matrix which is proven to have *almost surely* the smallest Frobenius loss asymptotically. It is compared with the linear shrinkage estimator constructed by Ledoit and Wolf [22]. A significant improvement is obtained when some additional *prior* information on the population covariance matrix is available.

Acknowledgments

The authors are grateful to the referee and the editor for their suggestions, which have improved the presentation in the paper. The first author is partly supported by the German Science Foundation (DFG) (DI1723/2-1) via the Research Unit 1735 "Structural Inference in Statistics: Adaptation and Efficiency". He also appreciates the financial support of the German Science Foundation (DFG) via the projects BO3521/2-2 and OK103/1-2 "Wishart Processes in Statistics and Econometrics: Theory and Applications".

Appendix

Here the proofs of the theorems are given.

Proof of Theorem 3.1. In order to prove Theorem 3.1 we use the MP equation (2.7). Before we proceed, we rewrite $\frac{1}{p}\mathrm{tr}(\mathbf{S}_n^2)$ for all $z \in \mathbb{C}^+$ in the following way

$$\frac{1}{p}\mathrm{tr}(\mathbf{S}_n^2) = -\frac{1}{2}\frac{\partial^2}{\partial z^2}\frac{1}{p}\frac{\mathrm{tr}(\mathbf{S}_n - 1/z\mathbf{I})^{-1}}{z}\bigg|_{z=0} = -\frac{1}{2}\frac{\partial^2}{\partial z^2}\left(\frac{m_{F_n}(1/z)}{z}\right)\bigg|_{z=0}, \tag{A.1}$$

where $m_{F_n}(1/z)$ is the Stieltjes transform introduced in (2.5). From Theorem 2.1 we know that $m_{F_n}(1/z)$ tends almost surely to a nonrandom $m_F(1/z)$ which satisfies the MP equation (2.7). Using this fact and denoting $\Theta_n(z) = \frac{m_{F_n}(1/z)}{z}$ we get that $\Theta_n(z)$ tends almost surely to $\Theta(z) = \frac{m_F(1/z)}{z}$ which is the unique solution of the following equation

$$\Theta(z) = \int_{-\infty}^{+\infty} \frac{1}{z\tau((1-c) - c\Theta(z)) - 1} dH(\tau). \tag{A.2}$$

Before we proceed with the second derivative of $\Theta(z)$ with respect to z, it is shown that the quantities $\Theta(z)$, $\Theta'(z)$ and $\Theta''(z)$, which appear in the calculation, are bounded at zero.

First, we point out that the limit under the integral sign in (A.2) can be safely moved by applying the dominated convergence theorem together with the fact that H is a probability measure and the boundedness of $\Theta(z)$. More precisely, let $\underline{m}(z) = -z(1-c) + cm_F(1/z)$ and rewrite (A.2) in the following way

$$\Theta(z) = -\int_{-\infty}^{+\infty} \frac{1}{\tau\underline{m}(z) + 1} dH(\tau), \tag{A.3}$$

where the function $\underline{m}(z)$ is another Stieltjes transformation of a positive measure on $\mathbb{R}^+$ (see, [26]). Thus, using inequality (see, [25, Lemma 6])

$$\left|\frac{\tau}{\tau\underline{m}(z) + 1}\right| \le \tau\frac{|z|}{\mathbf{Im}(z)} \tag{A.4}$$

we get

$$\left|\frac{1}{\tau\underline{m}(z) + 1}\right| \le \frac{|z|}{\mathbf{Im}(z)}. \tag{A.5}$$

Now, without loss of generality we construct the complex sequence $z_k = ih_k$ such that $h_k \to 0^+$ as $k \to \infty$. Note that for our framework the real part of z is not essential that is why without loss of generality we put it equal to zero. It follows that

$$\left|\frac{1}{\tau\underline{m}(z_k) + 1}\right| \le \frac{|z_k|}{\mathbf{Im}(z_k)} = \frac{h_k}{h_k} = 1. \tag{A.6}$$

So, using the fact that the function under the integral in (A.3) is bounded by an integrable function under the probability measure H and the dominated convergence theorem we can move the limit $z \to 0^+$ under the integral sign in (A.2). It leads to

$$\Theta(0) \equiv \lim_{z\to 0^+} \Theta(z) = -1, \quad \text{if } \lim_{z\to 0^+} z\Theta(z) < \infty. \tag{A.7}$$

The last inequality is true since if $\lim_{z\to 0^+} z\Theta(z) = \infty$ then from (A.2) we get that $\lim_{z\to 0^+} \Theta(z) = 0$. This leads to $\lim_{z\to 0^+} z\Theta(z) = 0$ which contradicts the statement $\lim_{z\to 0^+} z\Theta(z) = \infty$.

A similar analysis is performed for $\Theta'(z)$ given by

$$\Theta'(z) = -\int_{-\infty}^{+\infty} \frac{\tau((1-c) - c\Theta(z) - cz\Theta'(z))}{(z\tau((1-c) - c\Theta(z)) - 1)^2} dH(\tau). \tag{A.8}$$

Rearranging terms in (A.8) we get

$$\Theta'(z)\left(1 - cz\int_{-\infty}^{+\infty} \frac{\tau}{(z\tau((1-c) - c\Theta(z)) - 1)^2} dH(\tau)\right)$$
$$= \left(-(1-c) + c\Theta(z)\right)\int_{-\infty}^{+\infty} \frac{\tau}{(z\tau((1-c) - c\Theta(z)) - 1)^2} dH(\tau). \tag{A.9}$$

Due to $\Theta(0) < \infty$, we observe that the right hand side of (A.9) is bounded at zero and thus the left hand side tends to $\Theta'(0)$ as $z \to 0^+$. It leads to

$$\Theta'(0) \equiv \lim_{z\to 0^+} \Theta'(z) = -\int_{-\infty}^{+\infty} \tau dH(\tau). \tag{A.10}$$

Next, we derive $\Theta''(z)$. Let

$$M(z,\tau) = -\frac{1}{(z\tau((1-c) - c\Theta(z)) - 1)^2} \tag{A.11}$$

$$N(z,\tau) = \tau((1-c) - c\Theta(z) - cz\Theta'(z)). \tag{A.12}$$

Using (A.8) together with (A.11) and (A.12), $\Theta''(z)$ can be rewritten in the following way

$$\Theta''(z) = \int_{-\infty}^{+\infty} N'(z,\tau)M(z,\tau)dH(\tau) + \int_{-\infty}^{+\infty} M'(z,\tau)N(z,\tau)dH(\tau), \tag{A.13}$$

where

$$M'(z,\tau) = -2M(z,\tau)\frac{N(z,\tau)}{(z\tau((1-c) - c\Theta(z)) - 1)} \tag{A.14}$$

$$N'(z,\tau) = \tau\big(-2c\Theta'(z) - cz\Theta''(z)\big). \tag{A.15}$$

From (A.13) we get

$$\Theta''(z)\left(1 - cz\int_{-\infty}^{+\infty} \frac{\tau}{(z\tau((1-c) - c\Theta(z)) - 1)^2} dH(\tau)\right)$$
$$= 2c\Theta'(z)\int_{-\infty}^{+\infty} \frac{\tau}{(z\tau((1-c) - c\Theta(z)) - 1)^2} dH(\tau) + \int_{-\infty}^{+\infty} M'(z,\tau)N(z,\tau)dH(\tau). \tag{A.16}$$

Taking the limit as $z \to 0^+$ in (A.11), (A.12) and (A.14) as well as using (A.10) and (A.7) we obtain

$$\lim_{z\to 0^+} M(z,\tau) = -1, \tag{A.17}$$

$$\lim_{z\to 0^+} N(z,\tau) = \tau, \tag{A.18}$$

$$\lim_{z\to 0^+} M'(z,\tau) = -2\tau. \tag{A.19}$$

Hence, from (A.16) together with (A.17)–(A.19), (A.10) and (A.7) we get that $\Theta''(0) < \infty$ and

$$\Theta''(0) \equiv \lim_{z\to 0^+} \Theta''(z) = -2c\left(\int_{-\infty}^{+\infty} \tau dH(\tau)\right)^2 - 2\int_{-\infty}^{+\infty} \tau^2 dH(\tau). \tag{A.20}$$

Now the result of Theorem 3.1 follows from (A.1), (A.2) and (A.20). □

Proof of Theorem 3.2. In order to prove the statement of the Theorem 3.2 we use the following lemma of Rubio and Mestre [25].

Lemma A.1 (*Lemma 4, Rubio and Mestre [25]*)**.** *Let $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$ be a sequence of i.i.d. real random vectors with zero mean vector, identity covariance matrix, and uniformly bounded $4+\varepsilon$ moments for some $\varepsilon > 0$ and let $\mathbf{C}_n$ be some nonrandom matrix (possibly random but independent of $\boldsymbol{\xi}_n$) with bounded trace norm at infinity. Then*

$$\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i'\mathbf{C}_n\boldsymbol{\xi}_i - \operatorname{tr}(\mathbf{C}_n)\right| \longrightarrow 0 \quad \text{a.s. as } n \to \infty. \tag{A.21}$$

Next, we proceed to the proof of Theorem 3.2 directly by considering the asymptotic behavior of the following two quantities

$$\eta_1 = \frac{1}{p}\operatorname{tr}(\mathbf{S}_n\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'\left(\frac{1}{p}\boldsymbol{\Theta}^{1/2}\boldsymbol{\Sigma}_n\boldsymbol{\Theta}^{1/2}\right)\mathbf{x}_i \tag{A.22}$$

$$\eta_2 = \frac{1}{p}\|\mathbf{S}_n\|_F^2 = \frac{1}{p}\operatorname{tr}(\mathbf{S}_n^2), \tag{A.23}$$

where $\mathbf{x}_i$ is the ith column of the matrix $\mathbf{X}_n$ defined in (2.2).

First, we prove that for $\frac{p}{n} \to c \in (0, +\infty)$ as $n \to \infty$ the following assertion holds

$$\left|\eta_1 - \frac{1}{p}\operatorname{tr}(\boldsymbol{\Sigma}_n\boldsymbol{\Theta})\right| \longrightarrow 0 \quad \text{a.s. for } n \to \infty. \tag{A.24}$$

For the application of Lemma A.1 we have to show that the trace norm of the matrix $\frac{1}{p}\boldsymbol{\Theta}^{1/2}\boldsymbol{\Sigma}_n\boldsymbol{\Theta}^{1/2}$ is bounded. It holds that (see, [13])

$$\left\|\frac{1}{p}\boldsymbol{\Theta}^{1/2}\boldsymbol{\Sigma}_n\boldsymbol{\Theta}^{1/2}\right\|_{\text{tr}} = \frac{1}{p}\operatorname{tr}(\boldsymbol{\Theta}^{1/2}\boldsymbol{\Sigma}_n\boldsymbol{\Theta}^{1/2}) = \frac{1}{p}\operatorname{tr}(\boldsymbol{\Theta}\boldsymbol{\Sigma}_n) \le \frac{\operatorname{tr}(\boldsymbol{\Theta})\tau_{\max}(\boldsymbol{\Sigma}_n)}{p}, \tag{A.25}$$

where $\tau_{\max}(\boldsymbol{\Sigma}_n)$ denotes the largest eigenvalue of the matrix $\boldsymbol{\Sigma}_n$. At last, using the assumption (A5) we deduce the boundedness of the trace norm of the matrix $\frac{1}{p}\boldsymbol{\Theta}^{1/2}\boldsymbol{\Sigma}_n\boldsymbol{\Theta}^{1/2}$. Thus, using Lemma A.1 we get the statement (A.24).

Next, we prove the main result of Theorem 3.2, namely the following statement

$$\left|\eta_2 - \frac{1}{p}\left(\|\boldsymbol{\Sigma}_n\|_F^2 + \frac{c}{p}\|\boldsymbol{\Sigma}_n\|_{\text{tr}}^2\right)\right| \longrightarrow 0 \quad \text{a.s. for } \frac{p}{n} \to c \in (0, +\infty) \text{ as } n \to \infty. \tag{A.26}$$

In order to prove (A.26) we use the result of Theorem 3.1. First, we rewrite the difference in (A.26) via the triangle inequality in the following way

$$\left|\eta_2 - \frac{1}{p}\left(\|\boldsymbol{\Sigma}_n\|_F^2 + \frac{c}{p}\|\boldsymbol{\Sigma}_n\|_{\text{tr}}^2\right)\right| \le \left|\eta_2 - \phi\right| + \left|\phi - \frac{1}{p}\left(\|\boldsymbol{\Sigma}_n\|_F^2 + \frac{c}{p}\|\boldsymbol{\Sigma}_n\|_{\text{tr}}^2\right)\right|, \tag{A.27}$$

where $\phi = \int_{-\infty}^{+\infty}\tau^2 dH(\tau) + c\left(\int_{-\infty}^{+\infty}\tau dH(\tau)\right)^2$ is given in Theorem 3.1. Next we show that the right hand side of (A.27) vanishes almost surely as $n \to \infty$. Using Theorem 3.1 we get immediately that for the first term in (A.27) holds

$$\left|\eta_2 - \phi\right| \longrightarrow 0 \quad \text{a.s. for } n \to \infty. \tag{A.28}$$

The second term in (A.27) is nonrandom. Next, we show that it approaches to zero as $n \to \infty$. Due to assumption (A3) it holds that $H_n(t)$ tends to $H(t)$ at all continuity points of $H(t)$. Thus,

$$\frac{1}{p}\|\boldsymbol{\Sigma}_n\|_F^2 = \frac{1}{p}\operatorname{tr}(\boldsymbol{\Sigma}_n^2) = \frac{1}{p}\sum_{i=1}^{p}\tau_i^2 = \int_{-\infty}^{+\infty}\tau^2 dH_n(\tau) \xrightarrow{n\to\infty} \int_{-\infty}^{+\infty}\tau^2 dH(\tau). \tag{A.29}$$

The integral in (A.29) exists due to assumption (A5). Similarly, it can be shown that

$$\frac{1}{p^2}\|\boldsymbol{\Sigma}_n\|_{\text{tr}}^2 \xrightarrow{n\to\infty} \left(\int_{-\infty}^{+\infty}\tau dH(\tau)\right)^2. \tag{A.30}$$

From (A.29) and (A.30) it follows that

$$\left|\phi - \frac{1}{p}\left(\|\mathbf{\Sigma}_n\|_F^2 + \frac{c}{p}\|\mathbf{\Sigma}_n\|_{\mathrm{tr}}^2\right)\right| \longrightarrow 0 \quad \text{for } n \to \infty. \tag{A.31}$$

As a result, (A.28) and (A.31) complete the proof of Theorem 3.2. □

References

[1] J. Bai, Inferential theory for factor models of large dimensions, Econometrica 71 (2003) 135–171.
[2] Z.D. Bai, B.Q. Miao, G.M. Pan, On asymptotics of eigenvectors of large sample covariance matrix, Ann. Probab. 35 (4) (2007) 1532–1572.
[3] J. Bai, S. Ng, Determining the number of factors in approximate factor models, Econometrica 70 (2002) 191–221.
[4] J. Bai, S. Shi, Estimating high dimensional covariance matrices and its applications, Ann. Econom. Finance 12 (2) (2011) 199–215.
[5] Z.D. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, Springer, New York, Dordrecht, Heidelberg, London, 2010.
[6] P. Bühlmann, S. van de Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer, Heidelberg, New York, 2011.
[7] T. Cai, W. Lui, X. Luo, A constrained l_1 minimization approach to sparse precision matrix estimation, J. Amer. Statist. Assoc. 106 (2011) 594–607.
[8] T. Cai, X. Shen, High-Dimensional Data Analysis. Frontiers of Statistics: Volume 2, World Scientific, Singapore, 2011.
[9] T. Cai, M. Yuan, Adaptive covariance matrix estimation through block thresholding, Ann. Statist. 40 (2012) 2014–2042.
[10] T. Cai, H. Zhou, Minimax estimation of large covariance matrices under l_1 norm, Statist. Sinica 22 (2012) 1319–1378.
[11] J. Fan, Y. Fan, J. Lv, High dimensional covariance matrix estimation using factor model, J. Econometrics 147 (2008) 186–197.
[12] J. Fan, Y. Liao, M. Mincheva, Large covariance estimation by thresholding principal orthogonal complements, J. R. Stat. Soc. Ser. B 75 (2013) 603–680.
[13] Y. Fang, K.A. Loparo, X. Feng, Inequalities for the trace of matrix product, IEEE Trans. Automat. Control 39 (1994) 2489–2490.
[14] G. Frahm, C. Memmel, Dominating estimators for minimum-variance portfolios, J. Econometrics 159 (2010) 289–302.
[15] V.L. Girko, Theory of Random Determinants, Kluwer Academic Publishers, Dordrecht, 1990.
[16] V.L. Girko, Statistical Analysis of Observations of Increasing Dimension, Kluwer Academic Publishers, Dordrecht, Boston, London, 1995.
[17] V.L. Girko, A.K. Gupta, Asymptotic behaviour of spectral function of empirical covariance matrices, Random Oper. Stoch. Equ. 2 (1994) 43–60.
[18] V.L. Girko, A.K. Gupta, Canonical equation for the resolvent of empirical covariance matrices pencil, in: A.K. Gupta, V.L. Girko (Eds.), Multidimensional Statistical Analysis and Theory of Random Matrices, VSP Publishers, Netherlands, 1996.
[19] V.L. Girko, A.K. Gupta, Multivariate elliptically contoured linear models and some aspects of the theory of random matrices, in: A.K. Gupta, V.L. Girko (Eds.), Multidimensional Statistical Analysis and Theory of Random Matrices, VSP Publishers, Netherlands, 1996.
[20] G. Golub, C.F. Van Loan, Matrix Computations, third ed., The Johns Hopkins University Press, Baltimore, 1996.
[21] L. Le Cam, G. Lo Yang, Asymptotics in Statistics: Some Basic Concepts, Springer, New York, 2000.
[22] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2004) 365–411.
[23] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Sb. Math. 1 (1967) 457–483.
[24] A. Rohde, A.B. Tsybakov, Estimation of high dimensional low-rank matrices, Ann. Statist. 39 (2011) 887–930.
[25] F. Rubio, X. Mestre, Spectral convergence for a general class of random matrices, Statist. Probab. Lett. 81 (5) (2011) 592–602.
[26] J.W. Silverstein, Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices, J. Multivariate Anal. 55 (1995) 331–339.
[27] J.W. Silverstein, The Stieltjes transform and its role in eigenvalue behavior of large dimensional random matrices, in: Random Matrix Theory and its Applications, in: Lecture Notes Series, World Scientific, Singapore, 2009.
[28] J.W. Silverstein, Z.D. Bai, On the empirical distribution of eigenvalues of a class of large dimensional random matrices, J. Multivariate Anal. 54 (2) (1995) 175–192.
[29] J.W. Silverstein, S.-I. Choi, Analysis of the limiting spectral distribution of large-dimensional random matrices, J. Multivariate Anal. 54 (1995) 295–309.
[30] C. Stein, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I, Univ. California Press, Berkeley, 1956, pp. 197–206.
[31] Y.Q. Yin, Limiting spectral distribution for a class of random matrices, J. Multivariate Anal. 20 (1986) 50–68.