RCEA THE RIMINI CENTRE FOR ECONOMIC ANALYSIS

WP 10-17

Karim M. Abadir
Imperial College London
The Rimini Centre for Economic Analysis (RCEA)

Walter Distaso
Imperial College London

Filip Žikeš
Imperial College London

MODEL-FREE ESTIMATION OF LARGE VARIANCE MATRICES

Model-free estimation of large variance matrices[*]

Karim M. Abadir, Walter Distaso, Filip Žikeš

Imperial College London

May 4, 2010

Abstract

This paper introduces a new method for estimating large variance matrices. Starting from the orthogonal decomposition of the sample variance matrix, we exploit the fact that orthogonal matrices are never ill-conditioned and therefore focus on improving the estimation of the eigenvalues. We estimate the eigenvectors from just a fraction of the data, then use them to transform the data into approximately orthogonal series that we use to estimate a well-conditioned matrix of eigenvalues. Our estimator is model-free: we make no assumptions on the distribution of the random sample or on any parametric structure the variance matrix may have. By design, it delivers well-conditioned estimates regardless of the dimension of problem and the number of observations available. Simulation evidence show that the new estimator outperforms the usual sample variance matrix, not only by achieving a substantial improvement in the condition number (as expected), but also by much lower error norms that measure its deviation from the true variance.

Keywords.   variance matrices, ill-conditioning, mean squared error, mean absolute deviations, resampling.

JEL Classification: C10.

1 Introduction

Variance matrices are an important input for various applications in social sciences. Examples go from financial time series, where variance matrices are used as a measure of risk, to molecular biology, where they are used for gene classification purposes. Yet estimation of variance matrices is a statistically challenging problem, since the number of parameters grows as a quadratic function of the number of variables. To make things harder, conventional methods deliver nearly-singular (ill-conditioned) estimators when the dimension k of the matrix is large. As a result, estimators are very imprecise and operations such as matrix inversions amplify the estimation error further.

One strand of the literature has tackled this problem by trying to come up with methods that are able to achieve a dimensionality reduction by exploiting sparsity, imposing zero restrictions on some elements of the variance matrix. Wu and Pourahmadi (2003) and Bickel and Levina (2008a) propose banding methods to find consistent estimators of variance matrices (and their inverse). Other authors resort to thresholding (Bickel and Levina, 2008b, and El Karoui, 2009) or penalized likelihood methods (see, e.g., Fan and Peng, 2004 for the underlying general theory) to estimate sparse large variance matrices. Notable examples of papers using the latter method are Huang, Pourahmadi and Liu (2006) and Rothman, Bickel, Levina and Zhu (2008). Recently, Lam and Fan (2009) propose a unified theory of estimation, introducing the concept of *sparsistency*, which means that (asymptotically) the zero elements in the matrix are estimated as zero almost surely.

An alternative approach followed by the literature is to achieve dimensionality reduction using factor models. The idea is to replace the k individual series with a small number of unobservable factors such that they are able to capture most of the variation contained in the original data. Interesting examples are given by Fan, Fan and Lv (2008), Wang, Li, Zou and Yao (2009) and Lam and Yao (2009).

A third route is given by shrinkage, which entails substituting the original ill-conditioned estimator with a convex combination including it and a target matrix. The original idea is due to Stein (1956). Applications to variance matrix estimation include Jorion (1986), Muirhead (1987) and Ledoit and Wolf (2001, 2003, 2004). Intuitively, the role of the shrinkage parameter is to balance the estimation error coming from the ill-conditioned variance matrix and the specification error associated with the target matrix. Ledoit and Wolf (2001) propose an optimal estimation procedure for the shrinkage parameter, where the chosen metric is the Frobenius norm between the variance and the shrinkage matrix.

Finally, within the family of multivariate ARCH models, Engle, Shephard and Sheppard (2008) are able to estimate time-varying variance matrices allowing for the cross-sectional dimension to be larger than the time series one. The main idea behind their approach is to

use bivariate quasi-likelihoods for each pair of series and aggregate them into a composite likelihood. This also helps in improving the computational efficiency.

In this paper, we introduce a new method to estimate large nonsingular variance matrices. We propose a different approach for tackling this problem. Starting from the orthogonal decompositions of symmetric matrices, we exploit the fact that orthogonal matrices are never ill-conditioned, thus identifying the source of the problem as the eigenvalues. Our task is then to come up with an improved estimator of the eigenvalues. We achieve this by estimating the eigenvectors from just a fraction of the data, then using them to transform the data into approximately orthogonal series that we use to estimate a well-conditioned matrix of eigenvalues.

The result is a well-conditioned consistent estimator, which performs very well in terms of mean squared error and other traditional precision criteria. Because of the orthogonalization of the data, the resulting estimate is always nonsingular, even when the dimension of the matrix is larger than the sample size. Our estimator outperforms the traditional one, not only by achieving a substantial improvement in the condition number (as expected), but also by lower error norms that measure its deviation from the true variance. This is an important result, given that the existing literature shows that gains in reducing ill-conditioning are associated with small (or no) gains in the precision of the better-conditioned estimator (see, e.g., Fan, Fan and Lv, 2008).

Our method has a number of attractive features. First, it is model-free, in the sense that no assumptions are made on the densities of the random sample or on any underlying parametric model for the structure of the variance matrix. Second, it always delivers nonsingular well-conditioned estimators, hence remaining precise when further operations (such as inversions) are required.

This paper is organized as follows. Section 2 introduces the proposed estimator and establishes its main properties. By means of a Monte-Carlo experiment, Section 3 studies the finite-sample properties of the estimator and provides guidance on its use in practice. Finally, Section 4 concludes.

2 The new estimator

This section contains two parts. First, we briefly present the setup and the intuition for why the new estimator will perform well. Then, we investigate the estimator's properties and describe the optimal choice of two subsampling parameters. We do so for the simplest formulation of our estimator, which is to be generalized in Subsection 3.3 later.

2.1 The setup and the idea behind the estimator

Let $\boldsymbol{\Sigma} := \mathrm{var}(\boldsymbol{x})$ be a finite $k \times k$ positive definite variance matrix of $\boldsymbol{x}$. Suppose we have an i.i.d. sample $\{\boldsymbol{x}_i\}_{i=1}^n$, arranged into the $n \times k$ matrix $\boldsymbol{X} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ on which we base the usual estimator (ill-conditioned when k is large)

$$\widehat{\boldsymbol{\Sigma}} \equiv \widehat{\mathrm{var}}(\boldsymbol{x}) := \frac{1}{n} \boldsymbol{X}' \boldsymbol{M}_n \boldsymbol{X},$$

where $\boldsymbol{M}_n := \boldsymbol{I}_n - \frac{1}{n} \boldsymbol{\imath}_n \boldsymbol{\imath}_n'$ is the demeaning matrix of dimension n and $\boldsymbol{\imath}_n$ is a $n \times 1$ vector of ones. The assumption of an i.i.d. setup is not as restrictive as it may seem: the data can be filtered by an appropriate model (rather than just demeaning by $\boldsymbol{M}_n$) and the method applied to the residuals; for example, fitting a VAR model (if adequate) to a vector of time series and applying the method to the residuals. We will stick to the simplest setup, so as to clarify the workings of our method.

We can decompose this symmetric matrix as

(1) $$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{P}} \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{P}}',$$

where $\widehat{\boldsymbol{P}}$ is orthogonal and has typical column $\widehat{\boldsymbol{p}}_i$ ($i = 1, \ldots, k$), $\widehat{\boldsymbol{\Lambda}}$ being the diagonal matrix of eigenvalues of $\widehat{\boldsymbol{\Sigma}}$; e.g. see Abadir and Magnus (2005) for this and subsequent matrix results and notation in this paper. The condition number of any matrix is the ratio of the largest to smallest singular values of this matrix, a value of 1 being the best ratio. By orthogonality, all the eigenvalues of $\widehat{\boldsymbol{P}}$ lie on the unit circle and this matrix is always well-conditioned for any n and k. This leaves $\widehat{\boldsymbol{\Lambda}}$ as the source of the ill-conditioning of the estimate $\widehat{\boldsymbol{\Sigma}}$. We will therefore consider an improved estimator of $\boldsymbol{\Lambda}$: a simple consistent estimator of $\boldsymbol{P}$ will be used to transform the data to achieve approximate orthogonality of the transformed data (in variance terms), hence yielding a better-conditioned estimator of the variance matrix.

We can rewrite the decomposition as

(2) $$\widehat{\boldsymbol{\Lambda}} = \widehat{\boldsymbol{P}}' \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{P}} = \mathrm{diag}(\widehat{\mathrm{var}}(\widehat{\boldsymbol{p}}_1' \boldsymbol{x}), \ldots, \widehat{\mathrm{var}}(\widehat{\boldsymbol{p}}_k' \boldsymbol{x}))$$

since $\widehat{\boldsymbol{\Lambda}}$ is diagonal by definition. Now suppose that, instead of basing $\widehat{\boldsymbol{P}}$ on the whole sample, we base it on only m observations (say the first m ones, since the i.i.d. setup means that there is no gain from doing otherwise), use it to approximately orthogonalize the rest of the $n - m$ observations (as we did with $\widehat{\boldsymbol{p}}_i' \boldsymbol{x}$) which are then used to reestimate $\boldsymbol{\Lambda}$. Taking $m \to \infty$ and $n - m \to \infty$ as $n \to \infty$, standard statistical analysis implies that the resulting estimators are consistent. Notice that the choice of basing the second step on the remaining $n - m$ observations comes from two considerations. First, it is inefficient to discard observations in an i.i.d. setup, so we should not have fewer than these $n - m$

observations. Second, we should not reuse some of the first m observations because they worsen the estimate of $\boldsymbol{\Lambda}$, as will be seen in Proposition 1 below, hence making m the only subsampling parameter in question. Proposition 2 will show that the precision of the new estimator is optimized asymptotically by $m \propto n$ and will be followed by a discussion of how to calculate the optimal m by resampling in any finite sample.

2.2   Investigation of the estimator's properties

To summarize the procedure in equations, we start by writing

$$(3) \qquad \boldsymbol{X}' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) =: (\boldsymbol{X}_1', \boldsymbol{X}_2'),$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $m \times k$ and $(n-m) \times k$, respectively. Calculating $\widehat{\operatorname{var}}(\boldsymbol{x})$ based on the first m observations yields

$$(4) \qquad \frac{1}{m} \boldsymbol{X}_1' \boldsymbol{M}_m \boldsymbol{X}_1 = \widehat{\boldsymbol{P}}_1 \widehat{\boldsymbol{\Lambda}}_1 \widehat{\boldsymbol{P}}_1',$$

whence the desired first-step estimator $\widehat{\boldsymbol{P}}_1$. Then, estimate $\boldsymbol{\Lambda}$ from the remaining observations by

$$(5) \qquad \operatorname{dg}\left(\widehat{\operatorname{var}}(\widehat{\boldsymbol{P}}_1' \boldsymbol{x})\right) = \frac{1}{n-m} \operatorname{dg}\left(\widehat{\boldsymbol{P}}_1' \boldsymbol{X}_2' \boldsymbol{M}_{n-m} \boldsymbol{X}_2 \widehat{\boldsymbol{P}}_1\right) =: \widetilde{\boldsymbol{\Lambda}}$$

to replace $\widehat{\boldsymbol{\Lambda}}$ of (1) and obtain the new estimator

$$(6) \qquad \widetilde{\boldsymbol{\Sigma}} := \widehat{\boldsymbol{P}} \widetilde{\boldsymbol{\Lambda}} \widehat{\boldsymbol{P}}' = \frac{1}{n-m} \widehat{\boldsymbol{P}} \operatorname{dg}\left(\widehat{\boldsymbol{P}}_1' \boldsymbol{X}_2' \boldsymbol{M}_{n-m} \boldsymbol{X}_2 \widehat{\boldsymbol{P}}_1\right) \widehat{\boldsymbol{P}}'.$$

Note that we use the traditional estimator of variance matrices $\widehat{\operatorname{var}}(\cdot)$ in each of the two steps of our procedure, albeit in a different way. When we wish to stress the dependence of $\widetilde{\boldsymbol{\Sigma}}$ on the choice of m, we will write $\widetilde{\boldsymbol{\Sigma}}_m$ instead of $\widetilde{\boldsymbol{\Sigma}}$. There are three remarks to make here. First, by standard statistical analysis again, efficiency considerations imply that we should use $\operatorname{dg}(\widehat{\operatorname{var}}(\widehat{\boldsymbol{P}}_1' \boldsymbol{x}))$ rather than $\widehat{\operatorname{var}}(\widehat{\boldsymbol{P}}_1' \boldsymbol{x})$ in the second step given by (5)–(6), since by doing so we impose the correct restriction that estimators of $\boldsymbol{\Lambda}$ should be diagonal. Second, the estimate $\widetilde{\boldsymbol{\Sigma}}$ is almost surely nonsingular, like the true $\boldsymbol{\Sigma}$, because of the use of dg in (5). Third, we choose to demean $\boldsymbol{X}_2$ by its own mean (rather than the whole sample's mean) mainly for robustness considerations, in case the i.i.d. assumption is violated, e.g. due to a break in the *level* of the series.

We now turn to the issue of the choice of the last $n-m$ observations, rather than reusing some of the first m observations in addition to the last $n-m$ in (5). The following relies on asymptotic results, rather than the exact finite-sample arguments based on i.i.d. sampling that were used in the previous subsection.

Proposition 1 *Define $\boldsymbol{y}_i := \boldsymbol{x}_i - \overline{\boldsymbol{x}}$, where $\overline{\boldsymbol{x}} := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$, and consider the estimator*

$$\widetilde{\boldsymbol{\Lambda}}_j := \frac{1}{n-j} \, \mathrm{dg} \left(\widehat{\boldsymbol{P}}_1' \sum_{i=j+1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i' \widehat{\boldsymbol{P}}_1 \right)$$

for $j = 0, 1, \ldots, m$. It is assumed that the fourth-order moment of $\boldsymbol{x}$ exist. As $n - m \to \infty$ and $m \to \infty$, the condition number of $\widetilde{\boldsymbol{\Lambda}}_j$ is minimized with probability 1 by choosing $j/m \to 1$.

Before we prove this proposition, we make the following remark. The estimator $\widetilde{\boldsymbol{\Lambda}}_j$ differs slightly from the one used in (5) for $j = m$, because of the demeaning by the whole sample's mean $\overline{\boldsymbol{x}}$ in the proposition, as opposed to $\boldsymbol{X}_2' \boldsymbol{M}_{n-m} \boldsymbol{X}_2$ demeaning the last $n - m$ observations by their own sample mean. The difference tends to zero with probability 1 as $n - m \to \infty$ and does not affect the leading term of the expansions required in this proposition. Also, the assumption of the existence of the fourth-order moments for $\boldsymbol{x}$ is sufficient for the application of the limit theorem that we will use to prove the proposition, but it need not be a necessary condition.

Proof. For $m > j + 2$,

$$
\begin{aligned}
(7) \qquad \widetilde{\boldsymbol{\Lambda}}_j \;&=\; \frac{1}{n-j} \, \mathrm{dg} \left(\widehat{\boldsymbol{P}}_1' \sum_{i=j+1}^{m} \boldsymbol{y}_i \boldsymbol{y}_i' \widehat{\boldsymbol{P}}_1 \right) + \frac{1}{n-j} \, \mathrm{dg} \left(\widehat{\boldsymbol{P}}_1' \sum_{i=m+1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i' \widehat{\boldsymbol{P}}_1 \right) \\
&=\; \frac{m-j}{n-j} \, \mathrm{dg} \left(\boldsymbol{S}_j \right) + \frac{n-m}{n-j} \, \widetilde{\boldsymbol{\Lambda}}_m,
\end{aligned}
$$

which is a weighted average of $\mathrm{dg} \left(\boldsymbol{S}_j \right)$ and $\widetilde{\boldsymbol{\Lambda}}_m$ with

$$\boldsymbol{S}_j := \frac{1}{m-j} \widehat{\boldsymbol{P}}_1' \sum_{i=j+1}^{m} \boldsymbol{y}_i \boldsymbol{y}_i' \widehat{\boldsymbol{P}}_1.$$

Notice the special case $\boldsymbol{S}_0 = \widehat{\boldsymbol{\Lambda}}_1$ by (4), which is the ill-conditioned estimator that arises from the traditional approach. Intuitively, we should get a better-conditioned estimator here by giving more weight to the latter component of the weighted average, the one that $\widetilde{\boldsymbol{\Lambda}}_m$ represents. We will now show this by means of the law of iterated logarithm (LIL).

Recalling that $m, n - m \to \infty$ and $\widehat{\boldsymbol{P}}_1$ asymptotically orthogonalizes the two $\sum_i \boldsymbol{y}_i \boldsymbol{y}_i'$ sums in (7), the two limiting matrices for the components in (7) are both diagonal and we can omit the dg from $\boldsymbol{S}_j$. This omission is of order $1/\sqrt{m}$ and will not affect the optimization with respect to j, so we do not dwell on it in this proposition for the sake of clarity. It will however affect the optimization with respect to m, as we will see in the next proposition.

For any positive definite matrix, denote by λ_1 the largest and λ_k the smallest eigenvalue. The condition number is asymptotically equal to the ratio of the limsup to the liminf of the

diagonal elements (which are the eigenvalues here because of the diagonality of the limiting matrices) and is given with probability 1 by

$$c_n := \frac{\lambda_1 + \omega_1 \delta_n}{\lambda_k - \omega_k \delta_n},$$

where the LIL yields $\delta_n := \sqrt{2 \log{(\log{(n)})}/n}$ and ω_i^2/n as the asymptotic variance (which exists by assumption) of the estimator of λ_i. Writing c for $c_\infty = \lambda_1/\lambda_k$,

$$(8) \qquad c_n = \frac{\lambda_1 + \omega_1 \delta_n}{\lambda_k - \omega_k \delta_n} = \left(c + \frac{\omega_1 \delta_n}{\lambda_k} \right) \left(1 + \frac{\omega_k \delta_n}{\lambda_k} + O(\delta_n^2) \right) = c + \frac{\omega_1 + c \omega_k}{\lambda_k} \delta_n + O(\delta_n^2).$$

This last expansion is not necessary to establish our result, but it will clarify the objective function. Applying this formula to the two matrices in (7) and dropping the remainder terms, we get the asymptotic condition number of $\widetilde{\Lambda}_j$ as

$$C := c + \frac{\omega_1 + c \omega_k}{\lambda_k (n - j)} \left(\sqrt{2 (m - j) \log{(\log{(m - j)})}} + \sqrt{2 (n - m) \log{(\log{(n - m)})}} \right),$$

which is minimized by letting $j/m \to 1$ since $\lim_{a \to 0} a \log{(\log a)} = 0$ and $n > m$ (hence $n - j \geq 1$). The condition $m > j + 2$, given at the start of the proof, ensures that $\log{(m - j)} > 1$ and that C is real. The cases $m = j, j + 1, j + 2$ are not covered separately in this proof, because they are asymptotically equivalent to $m = j + 3$ as $m \to \infty$. $\blacksquare$

Note the conditions $n - m \to \infty$ and $m \to \infty$, needed for the consistency of the estimator. We now turn to the final question, inquiring how large m should be, relative to n. As in the previous proposition, the approach will be asymptotic. We will need, in addition, to assume the existence of fourth-order moments for $\boldsymbol{x}$ when we consider MSE-type criteria for the estimation of $\boldsymbol{\Sigma}$.

Define the following criteria for the (inverse) precision of the new estimator $\widetilde{\boldsymbol{\Sigma}}$:

$$(9) \qquad\qquad R_l(\widetilde{\boldsymbol{\Sigma}}) := E(|| \operatorname{vec}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) ||_l^l), \qquad l = 1, 2,$$

and

$$(10) \qquad\qquad R_{l,S}(\widetilde{\boldsymbol{\Sigma}}) := E(|| \operatorname{vech}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}) ||_l^l), \qquad l = 1, 2,$$

where the l-th norm is $||\boldsymbol{a}||_l := (\sum_{i=1}^j |a_i|^l)^{1/l}$ for any j-dimensional vector $\boldsymbol{a}$. In the case of $k = 1$, these criteria reduce to the familiar mean absolute deviation (MAD) and mean squared error (MSE) for $l = 1, 2$, respectively. The half-vec operator, vech, selects only the distinct elements of a general symmetric matrix. If $\boldsymbol{\Sigma}$ has a Toeplitz structure (as in the case of autocovariance matrices for stationary series), σ_{ij} depends only on the distance $|i - j|$ from the diagonal and we define

$$(11) \qquad\qquad R_{l,T}(\widetilde{\boldsymbol{\Sigma}}) := E(||(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{e}_1||_l^l), \qquad l = 1, 2,$$

where $\boldsymbol{e}_1 := (1, \boldsymbol{0}'_{k-1})'$ is the first elementary vector: postmultiplying a matrix with it selects the first column of that matrix.

Since the dimension k is finite as n increases, there is asymptotically no difference in considering the usual, the S, or the T version for each l. However, we advocate the use of the relevant criterion in finite samples, in order to give the same weight to each distinct element in $\boldsymbol{\Sigma}$.

Proposition 2 *As $n - m \to \infty$ and $m \to \infty$, the precision criteria in (9)–(11) are optimized asymptotically by taking $m \propto n$.*

Proof. The result will be obtained by deriving a stochastic expansion for $\widetilde{\boldsymbol{\Lambda}}$ and balancing the nonzero second-order terms in m and n.

Since standard asymptotics give that $\widehat{\boldsymbol{\Sigma}}_1 = \widehat{\boldsymbol{P}}_1 \widehat{\boldsymbol{\Lambda}}_1 \widehat{\boldsymbol{P}}'_1$ is $\sqrt{m}$-consistent, both components $\widehat{\boldsymbol{P}}_1, \widehat{\boldsymbol{\Lambda}}_1$ will be $\sqrt{m}$-consistent:

$$\widehat{\boldsymbol{P}}_1 = \boldsymbol{P}\left(1 + O_p\left(1/\sqrt{m}\right)\right) \qquad \text{and} \qquad \widehat{\boldsymbol{\Lambda}}_1 = \boldsymbol{\Lambda}\left(1 + O_p\left(1/\sqrt{m}\right)\right);$$

otherwise, $\widehat{\boldsymbol{\Sigma}}_1 \neq \boldsymbol{\Sigma}\left(1 + O_p\left(1/\sqrt{m}\right)\right)$ in general. Furthermore, the $O_p\left(1/\sqrt{m}\right)$ terms will be nonzero because of the standard i.i.d. setup. Defining $\boldsymbol{y}_i := \boldsymbol{x}_i - \overline{\boldsymbol{x}}_2$, where $\overline{\boldsymbol{x}}_2 := \frac{1}{n-m}\sum_{i=m+1}^{n} \boldsymbol{x}_i$, we have

$$\widetilde{\boldsymbol{\Lambda}} = \frac{1}{n-m} \operatorname{dg}\left(\widehat{\boldsymbol{P}}'_1 \sum_{i=m+1}^{n} \boldsymbol{y}_i \boldsymbol{y}'_i \widehat{\boldsymbol{P}}_1\right) = \operatorname{dg}\left(\widehat{\boldsymbol{P}}'_1 \widehat{\boldsymbol{\Sigma}}_2 \widehat{\boldsymbol{P}}_1\right),$$

where $\widehat{\boldsymbol{\Sigma}}_2 = \boldsymbol{\Sigma}\left(1 + O_p\left(1/\sqrt{n-m}\right)\right)$ and the $O_p\left(1/\sqrt{n-m}\right)$ term is nonzero. As a result,

$$\begin{aligned}\widetilde{\boldsymbol{\Lambda}} &= \operatorname{dg}\left(\boldsymbol{P}'\left(1 + O_p\left(1/\sqrt{m}\right)\right)\boldsymbol{\Sigma}\left(1 + O_p\left(1/\sqrt{n-m}\right)\right)\boldsymbol{P}\left(1 + O_p\left(1/\sqrt{m}\right)\right)\right)\\ &= \boldsymbol{\Lambda} + O_p\left(1/\sqrt{m}\right) + O_p\left(1/\sqrt{n-m}\right) + O_p\left(1/\sqrt{m(n-m)}\right).\end{aligned}$$

Balancing the remainder terms gives $m \propto n$: the term with a larger order of magnitude dominates asymptotically and should be reduced until all three terms are equally small, whatever the chosen norm ($l = 1, 2$). The stated result follows from the definition of the estimator $\widetilde{\boldsymbol{\Sigma}}$ in (6) and $\widehat{\boldsymbol{P}}$ being unaffected by m for any given n. ∎

Because of the i.i.d. setup, we can use resampling methods as a means of automation of the choice of m for any sample size. Standard proofs of the validity of such procedures apply here too. We shall illustrate with the bootstrap in the next section.

3   Simulations

In this section, we run a Monte Carlo experiment to study the finite-sample properties of our two-step estimator and compare it to the usual sample variance matrix. In particular,

we are interested in answering the question of how large m should be relative to n in order to balance the estimation of $\boldsymbol{P}$ (need large m) and the estimation of $\boldsymbol{\Lambda}$ (need small m). To this end, we employ the precision criteria defined in equations (9)–(11). Additionally, we investigate the reduction in the condition number of the two-step estimator $(\widetilde{c}_{n-m})$ relative to the sample variance matrix benchmark $(\widehat{c}_n)$. We do these calculations both for the simple version of our estimator seen so far, as well as for a more elaborate one introduced in Subsection 3.3 where we investigate its performance. Finally, Subsection 3.4 investigates the automation of the choice of m.

Our simulation design is as follows. We let the true variance matrix $\boldsymbol{\Sigma}$ have Toeplitz structure with typical element given by $\sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \ldots, k$, where $\rho \in [0, 1]$ and $k \in \{30, 100, 250\}$ is the dimension of the random vector $\boldsymbol{x}$ to which the variance matrix pertains. We consider three different correlation values, $\rho \in \{0.5, 0.75, 0.95\}$, covering scenarios of mild ($\rho = 0.5$) and relatively high ($\rho = 0.75, 0.95$) correlation but omitting the case $\rho = 0$ of a scalar matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_k$.[1] The random vector $\boldsymbol{x}$ is drawn either from the normal distribution or from the multivariate Student t distribution with eight degrees of freedom (denoted t(8)), with population mean equal to zero (without loss of generality). All simulations are based on 1,000 Monte Carlo replications, to save computational time. For example, the simulations for $k = 250$ and $n = 500$ with 1,000 replications required about 73 hours of run time. The exercise for $k = 30$ was repeated with 10,000 replications and essentially identical results were obtained.

3.1   Estimator's Precision

We start with dimension $k = 30$ and simulate random samples of size $n \in \{20, 30, 50\}$. The results are summarized in the left panels of Tables 1–2. Starting from the mean squared error, $\mathrm{R}_{2,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}})$, reported in Table 1, we see that the two-step estimator is more precise in small samples as long as ρ is less than the extreme 0.95. The MSE tends to be smaller for small values of m. The reduction in the mean squared error is very large, but decreases with ρ and is more pronounced for data generated from the fat-tailed Student t distribution.

Restricting attention to the distinct elements of the true variance matrix by looking at $\mathrm{R}_{2,\mathrm{T}}(\widetilde{\boldsymbol{\Sigma}})$ in Table 2, we get similar conclusions: large gains in precision are found for small values of m, except for the case of $\rho = 0.95$. However, we find that the optimal m is slightly larger in the case of $\mathrm{R}_{2,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}})$ than for $\mathrm{R}_{2,\mathrm{T}}(\widetilde{\boldsymbol{\Sigma}})$. This is due to the former criterion giving

---

[1]The case of a scalar matrix is obvious (and differs from the slightly more general case of diagonal matrices which we simulated for a few cases and will report qualitatively). This is because $\boldsymbol{Q}(\sigma^2 \boldsymbol{I}_k)\boldsymbol{Q}' = \sigma^2 \boldsymbol{I}_k$ for any orthogonal $\boldsymbol{Q}$, not necessarily the true $\boldsymbol{P}$ that we try to estimate. In other words, we do not need consistency of the estimate of $\boldsymbol{P}$ as a result, and the optimal choice of m is as small as possible. The case of a scalar matrix (which is rare) is essentially the case of scalar estimation of one variance.

more weight to repeated off-diagonal elements which we have seen to require a larger m. The choice of criterion to use in practice will depend on the user's objective function: some applications may prefer the precision of all the elements (e.g. for CAPM estimation in finance), while others may care only about the distinct underlying parameters of interest.

The results for $k = 100$ and $k = 250$ are reported in the left panels of Tables 3–4 and 5–6, respectively. In the case of $k = 100$ we simulate samples of size $n \in \{50, 100, 250\}$, whereas, for $k = 250$, $n \in \{100, 250, 500\}$.

Across the various precision measures, we observe similar patterns to the case of $k = 30$. Sizeable improvements in precision are found for all data generating processes except $\rho = 0.95$. For the alternative precision measures, $R_2(\widetilde{\Sigma})$, $R_1(\widetilde{\Sigma})$, $R_{1,S}(\widetilde{\Sigma})$, and $R_{1,T}(\widetilde{\Sigma})$, the results are qualitatively similar and are omitted to save space. The main difference is that the optimal m for the MAD criteria are determined largely by n, and are robust to the dimensions k, to the distribution (Gaussian or t(8)), and to ρ as long as it was not the extreme $\rho = 0.95$. These m were comparable to the MSE-optimal ones for intermediate values of ρ, but holding across the range, hence somewhat smaller overall.

3.2 Reduction in ill-conditioning

Moving to analyze the reduction in ill-conditioning, the left panel of Table 7 reports the average ratio of condition numbers $\widetilde{c}_{n-m}/\widehat{c}_n$ for k, n and m. Note that for $n \leq k$, the sample variance matrix is singular and hence its condition number is not defined. Starting from the case $k = 30$, we find that choosing small m delivers the largest improvements in the conditioning of the estimated variance matrix, and that the gains are massive. Moreover, the condition number appears to be relatively insensitive to the choice of m as long as it remains smaller than approximately half of the sample size. Within this range, the two-step estimator achieves about 15-20 times smaller condition number than the sample variance matrix. See also Figure 1, for a graphical display of the results.

The general picture emerging from the case $k = 30$ is confirmed for $k = 100$ and $k = 250$, with gains that are generally increasing in k and condition numbers that seem to be less sensitive to m as k increases.

An attractive feature of the two step estimator is that the reduction in ill-conditioning is preserved even in situations where $k \geq n$ and the conventional estimator is not positive definite. Unreported simulation results show that, for example, when $n = 20$, $k = 30$, $m = 5$, condition numbers for $\widetilde{\Sigma}$ are on average 40% higher than the corresponding ones obtained when $n = 50$, $k = 30$, $m = 5$, but still much lower than those of the sample variance matrix.

3.3 Resampling and averaging

In the previous section, we constructed the estimator $\widetilde{\boldsymbol{\Sigma}}$ by basing $\widehat{\boldsymbol{P}}_1$ on the first m observations in the sample and using it to approximately orthogonalize the remaining $n - m$ observations. This is, of course, only one out of $\binom{n}{m}$ possibilities of choosing the m observations to calculate $\widehat{\boldsymbol{P}}_1$, all of them being equivalent due to the i.i.d. assumption. In this subsection, we investigate how much does averaging over the different possibilities improve our original estimator $\widetilde{\boldsymbol{\Sigma}}$. The intuition behind this is that averaging will reduce the variability that comes with the choice of any one specific combination of m observations. More specifically, averaging over $\widetilde{\boldsymbol{\Sigma}}_m := \widehat{\boldsymbol{P}} \widetilde{\boldsymbol{\Lambda}}_m \widehat{\boldsymbol{P}}'$ is the same as averaging over $\widetilde{\boldsymbol{\Lambda}}_m$, since $\widehat{\boldsymbol{P}}$ is based on the full sample and hence does not depend on m. As a result, the asymptotic expansions of $\widetilde{\boldsymbol{\Lambda}}_m$ (e.g. as in Proposition 2) lose the higher-order terms that have zero mean.

Let $\boldsymbol{X}'_s := (\boldsymbol{X}'_{1,s}, \boldsymbol{X}'_{2,s})$, where $\boldsymbol{X}'_{1,s} := (\boldsymbol{x}^s_{1,1}, \ldots, \boldsymbol{x}^s_{1,m})$ is obtained by randomly sampling without replacement m columns from the original $\boldsymbol{X}'$, and $\boldsymbol{X}'_{2,s} := (\boldsymbol{x}^s_{2,1}, \ldots, \boldsymbol{x}^s_{2,n-m})$ is filled up with the remaining $n - m$ columns from $\boldsymbol{X}'$ that were not selected for $\boldsymbol{X}_{1,s}$. Let $\widetilde{\boldsymbol{\Sigma}}_{m,s}$ denote our estimator calculated from $\boldsymbol{X}_s$, and let

$$(12) \qquad \widetilde{\boldsymbol{\Sigma}}_{m,S} := \frac{1}{S} \sum_{s=1}^{S} \widetilde{\boldsymbol{\Sigma}}_{m,s}$$

denote the average over S samples. Computational burden makes the choice $S = \binom{n}{m}$ prohibitive, especially for large n. Unreported simulation results nonetheless show that relatively small number of samples S suffices to reap most of the benefits of averaging. To illustrate them, we have included Figure 2, where we vary S on the horizontal axis: we can see that all the benefits to be achieved occur very quickly for small S and there is not much to be gained from taking a large S. We simulated $S = 10, 20, \ldots, 100$ but only plotted $10, \ldots, 50$ together with no-averaging at the origin of the horizontal axis.

The right panels of Tables 1–7 report results obtained by averaging over $S = 20$ samples. Starting from efficiency measures, one can immediately gauge the improvements due to averaging. For $k = 30$ (Tables 1 and 2), the various efficiency measures are improved markedly. Benefits decrease as ρ increases, but remain substantial (about 50%).[2] For $k = 100$ (Tables 3 and 4), we observe a similar pattern. In the case of the student t distribution, the two-step estimator is now more (or as) efficient than the sample variance matrix even when $\rho = 0.95$. For large matrices ($k = 250$, Tables 5 and 6), we observe that the the two-step estimator is now more efficient than the sample variance matrix when $\rho = 0.95$ in the Gaussian case as well. Efficiency improvements are typically larger than

---

[2]Unreported simulation results show that, in the case of diagonal matrices, averaging improves efficiency measures by an order of magnitude.

those obtained by other approaches to estimate variance matrices (shrinkage, banding and thresholding; see Bickel and Levina, 2008b, Table 1, p.2598).

Important improvements are also observed in the condition number of the averaged two-step estimator (see Table 7). The largest (relative) improvements are obtained for $k = 30$, are slightly decreasing in k and seem to be uniform over the different values of ρ.

3.4 Data-dependent procedures to choose m

We next turn to the optimal choice of m in practical applications. One possibility is to use resampling techniques. The i.i.d. setup of the previous section implies that standard bootstrap applies directly to our estimator. We denote by $\boldsymbol{X}_b := \left(\boldsymbol{x}_1^b, \ldots, \boldsymbol{x}_n^b\right)'$ a bootstrap sample obtained by drawing independently n observations with replacement from the original sample $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$. The corresponding bootstrap versions of $\widehat{\boldsymbol{\Sigma}}$ and $\widetilde{\boldsymbol{\Sigma}}_m$ are denoted by $\widehat{\boldsymbol{\Sigma}}_b$ and $\widetilde{\boldsymbol{\Sigma}}_{m,b}$, respectively. Given B independent replications of $\widehat{\boldsymbol{\Sigma}}_b$ and $\widetilde{\boldsymbol{\Sigma}}_{m,b}$, we define

$$\widehat{\boldsymbol{\Sigma}}_B := \frac{n}{(n-1)B} \sum_{b=1}^{B} \widehat{\boldsymbol{\Sigma}}_b, \quad \text{and} \quad \widetilde{\boldsymbol{\Sigma}}_{m,B} := \frac{1}{B} \sum_{b=1}^{B} \widetilde{\boldsymbol{\Sigma}}_{m,b},$$

where $\widehat{\boldsymbol{\Sigma}}_B$ is the average bootstrapped sample variance matrix rescaled in order to remove the bias (which is $O(1/n)$), and $\widetilde{\boldsymbol{\Sigma}}_{m,B}$ is the average bootstrapped $\widetilde{\boldsymbol{\Sigma}}_m$. To balance the trade-off between variance and bias, we find the m that minimizes

$$\frac{1}{B} \sum_{b=1}^{B} (\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widetilde{\boldsymbol{\Sigma}}_{m,B}))'(\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widetilde{\boldsymbol{\Sigma}}_{m,B}))$$

$$+ \left(\frac{1}{B} \sum_{b=1}^{B} \text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widehat{\boldsymbol{\Sigma}}_B) \right)' \left(\frac{1}{B} \sum_{b=1}^{B} \text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widehat{\boldsymbol{\Sigma}}_B) \right),$$

where the first term estimates the "variance" associated with the distinct elements of $\widetilde{\boldsymbol{\Sigma}}_m$, while the second term approximates the squared "bias". Simple algebra shows that minimizing this objective function with respect to m is equivalent to minimizing

$$\frac{1}{B} \sum_{b=1}^{B} ||\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widehat{\boldsymbol{\Sigma}}_B)||_2^2,$$

which is computationally more convenient (it is also possible to use the l_1 norm instead of l_2 norm). In practice, we set up a grid $\mathcal{M} = \{m_1, m_2, \ldots, m_M\}$, where $1 \leq m_1 \leq m_M \leq n$ and calculate the objective function for each $m \in \mathcal{M}$. The grid may be coarser or finer depending on the available computational resources. The bootstrap-based optimal m is then given by

$$m_B := \underset{m \in \mathcal{M}}{\text{argmin}} \frac{1}{B} \sum_{b=1}^{B} ||\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widehat{\boldsymbol{\Sigma}}_B)||_2^2.$$

Results are reported in Table 8 and in general show a very good performance of the suggested procedure. The bootstrap selects m_B very close to the optimal m and the percentage increase in the mean squared error of the bootstrap-based estimator is minimal, well below 1%.

Another possibility is offered by observing that the precision of $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ is relatively stable over a wide range of m, approximately $m \in [0.2n, 0.8n]$ (see Tables 1–6). This suggests to construct an estimator based on averaging $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ over m. This "grand average" estimator is defined as

$$(13) \qquad \widetilde{\boldsymbol{\Sigma}}_{M,S} := \frac{1}{M} \sum_{m \in \mathcal{M}} \widetilde{\boldsymbol{\Sigma}}_{m,S},$$

where M is the number of elements in $\mathcal{M}$. Results are reported in Tables 1–7. The performance of $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ is very good in terms of precision. In many cases $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ is the most precise estimator. The price to pay for this increase in precision is a slight increase in ill-conditioning (with respect to $\widetilde{\boldsymbol{\Sigma}}$ calculated with the smallest m), since the condition number seems to be monotonically increasing in m.

4 Conclusion

In this paper, we provide a novel approach to estimate large variance matrices. Exploiting the properties of symmetric matrices, we are able to identify the source of ill-conditioning related to the standard sample variance matrix and hence provide an improved estimator. Our approach delivers more precise and well-conditioned estimators, regardless of the dimension of the problem and of the sample size. Theoretical findings are confirmed by the results of a Monte-Carlo experiment, which also offers some guidance on how to use the estimator in practice.

The substantial reduction in ill-conditioning suggests that our estimator should perform well in cases where matrix inversions operations are required, as for example in portfolio optimization problems. This substantial application is currently being investigated and is beyond the scope of this paper.

Table 1: $R_{2,S}(\widetilde{\boldsymbol{\Sigma}})$ for $k = 30$. [a]

		No averaging						Averaging over 20 samples					
		Gaussian			t(8)			Gaussian			t(8)		
		ρ			ρ			ρ			ρ		
n	m	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95
20	5	11.1	25.8	53.7	13.4	29.8	73.8	8.57	21.2	40.1	9.61	23.7	53.9
	10	11.5	23.5	54.0	15.1	30.3	83.3	8.20	18.1	37.2	9.35	21.3	53.5
	15	14.8	29.0	84.1	21.4	39.0	130	9.37	17.8	39.8	10.5	23.2	54.7
	av	-	-	-	-	-	-	8.19	18.4	37.1	9.19	21.2	50.9
	$\widehat{\boldsymbol{\Sigma}}$	**24.1**	**25.3**	**32.8**	**37.5**	**39.3**	**54.0**	**24.1**	**25.3**	**32.8**	**37.5**	**39.3**	**54.0**
30	5	9.88	22.6	43.8	11.3	25.6	56.1	8.10	18.9	29.9	8.86	20.9	39.5
	10	9.17	17.8	34.0	11.9	22.2	50.9	7.19	14.3	25.6	8.08	16.9	37.0
	15	9.30	17.5	36.5	11.7	23.0	56.2	6.94	13.6	24.8	7.93	16.5	36.8
	20	10.3	19.3	47.3	13.7	26.3	71.9	7.17	14.1	24.9	8.22	17.2	36.7
	25	13.9	26.3	82.6	18.7	36.6	113	8.67	16.3	29.8	9.73	19.5	41.1
	av	-	-	-	-	-	-	6.93	13.7	24.5	7.86	16.4	36.1
	$\widehat{\boldsymbol{\Sigma}}$	**16.3**	**17.1**	**22.5**	**24.7**	**26.6**	**36.3**	**16.3**	**17.1**	**22.5**	**24.7**	**26.6**	**36.3**
50	5	8.95	19.9	35.3	9.86	21.9	44.1	7.66	16.7	21.9	8.25	18.3	28.3
	10	7.66	13.7	22.1	8.83	16.9	32.3	6.35	11.0	16.5	7.10	13.3	23.8
	20	6.87	11.6	19.6	8.53	15.8	32.1	5.47	9.16	15.0	6.41	11.8	22.6
	30	7.22	12.2	25.4	9.44	17.3	41.1	5.41	9.31	14.9	6.47	12.1	22.5
	40	9.12	16.0	44.4	12.7	23.8	69.2	6.03	10.3	16.4	7.16	13.3	24.2
	av	-	-	-	-	-	-	5.37	9.08	14.6	6.30	11.6	21.4
	$\widehat{\boldsymbol{\Sigma}}$	**9.88**	**10.3**	**13.6**	**15.1**	**16.3**	**21.7**	**9.88**	**10.3**	**13.6**	**15.1**	**16.3**	**21.7**

[a] Notes: This Table reports $R_{2,S}(\widetilde{\boldsymbol{\Sigma}})$, averaged over the number of simulations. Results are given for different values of m, n and ρ, and for different multivariate distributions with mean zero and Toeplitz variance matrix $\boldsymbol{\Sigma}$ with typical element given by $\sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \ldots, k$. Bold entries refer to the sample variance matrix ($n = m$) and shaded cells report the minimum value over m. The left panel reports entries for the case where the first m observations in the sample are used to calculate $\widehat{\boldsymbol{P}}_1$ and the remaining ones are used to calculate $\widehat{\boldsymbol{\Lambda}}$. The right panel reports entries for the case where we randomly sample the m observations $S = 20$ times and average the resulting estimator to obtain $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ as in (12). The line "av" reports entries for the case where we average the estimator over different values of m, namely $m \in [0.2n, 0.8n]$, and obtain $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ as in (13). All results are based on 1,000 simulations.

Table 2: $R_{2,\mathrm{T}}(\widetilde{\boldsymbol{\Sigma}})$ for $k=30$.[a]

		No averaging						Averaging over 20 samples					
		Gaussian			t(8)			Gaussian			t(8)		
		ρ			ρ			ρ			ρ		
n	m	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95
	5	0.47	1.16	3.13	0.58	1.35	4.28	0.32	0.89	2.11	0.36	0.99	2.84
$n=20$	10	0.52	1.14	3.02	0.70	1.47	4.67	0.32	0.80	2.05	0.36	0.95	2.94
	15	0.70	1.41	4.32	1.07	2.03	6.81	0.38	0.89	2.12	0.42	1.06	2.93
	av	-	-	-	-	-	-	0.31	0.78	2.00	0.35	0.91	2.71
	$\widehat{\boldsymbol{\Sigma}}$	**1.49**	**1.54**	**1.94**	**2.28**	**2.31**	**3.10**	**1.49**	**1.54**	**1.94**	**2.28**	**2.31**	**3.10**
	5	0.42	1.03	2.62	0.48	1.15	3.25	0.31	0.80	1.57	0.33	0.88	2.08
	10	0.41	0.87	1.99	0.50	1.08	2.91	0.28	0.63	1.43	0.31	0.75	2.07
$n=30$	15	0.43	0.89	2.03	0.54	1.16	3.04	0.28	0.64	1.40	0.32	0.78	2.09
	20	0.50	0.99	2.50	0.67	1.36	3.74	0.31	0.69	1.39	0.35	0.84	2.07
	25	0.69	1.29	4.14	0.92	1.81	5.62	0.37	0.77	1.55	0.41	0.94	2.17
	av	-	-	-	-	-	-	0.28	0.63	1.37	0.31	0.75	1.96
	$\widehat{\boldsymbol{\Sigma}}$	**1.02**	**1.04**	**1.32**	**1.54**	**1.57**	**2.06**	**1.02**	**1.04**	**1.32**	**1.54**	**1.57**	**2.06**
	5	0.37	0.91	2.12	0.41	1.00	2.69	0.29	0.72	1.11	0.31	0.78	1.46
	10	0.33	0.66	1.34	0.38	0.82	1.96	0.25	0.48	0.91	0.28	0.57	1.32
$n=50$	20	0.32	0.61	1.10	0.39	0.85	1.83	0.23	0.45	0.85	0.27	0.57	1.28
	30	0.36	0.66	1.35	0.46	0.93	2.24	0.25	0.49	0.84	0.29	0.62	1.27
	40	0.47	0.84	2.22	0.65	1.23	3.55	0.28	0.53	0.89	0.33	0.68	1.33
	av	-	-	-	-	-	-	0.23	0.45	0.83	0.26	0.57	1.22
	$\widehat{\boldsymbol{\Sigma}}$	**0.61**	**0.63**	**0.77**	**0.93**	**0.97**	**1.26**	**0.61**	**0.63**	**0.77**	**0.93**	**0.97**	**1.26**

[a] Notes: This Table reports $R_{2,\mathrm{T}}(\widetilde{\boldsymbol{\Sigma}})$, averaged over the number of simulations. Results are given for different values of m, n and ρ, and for different multivariate distributions with mean zero and Toeplitz variance matrix $\boldsymbol{\Sigma}$ with typical element given by $\sigma_{ij} = \rho^{|i-j|}$, $i,j = 1,\ldots,k$. Bold entries refer to the sample variance matrix ($n=m$) and shaded cells report the minimum value over m. The left panel reports entries for the case where the first m observations in the sample are used to calculate $\widehat{\boldsymbol{P}}_1$ and the remaining ones are used to calculate $\widetilde{\boldsymbol{\Lambda}}$. The right panel reports entries for the case where we randomly sample the m observations $S=20$ times and average the resulting estimator to obtain $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ as in (12). The line "av" reports entries for the case where we average the estimator over different values of m, namely $m \in [0.2n, 0.8n]$, and obtain $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ as in (13). All results are based on 1,000 simulations.

Table 3: $R_{2,S}(\widetilde{\boldsymbol{\Sigma}})$ for $k = 100$.[a]

		No averaging						Averaging over 20 samples					
		Gaussian			t(8)			Gaussian			t(8)		
		ρ			ρ			ρ			ρ		
n	m	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95
	5	33.4	102	272	36.2	107	311	30.7	96.5	216	32.1	100	246
	10	32.3	85.7	180	35.4	92.7	235	29.1	79.0	138	30.8	84.5	177
50	20	31.2	71.1	155	35.4	81.9	222	27.2	63.6	124	29.1	70.9	166
	30	32.0	69.7	169	38.1	84.0	244	25.6	59.8	126	28.6	67.9	168
	40	37.5	79.1	235	47.6	101	330	27.7	62.2	132	29.7	70.6	184
	av	-	-	-	-	-	-	26.6	62.2	122	28.5	69.6	163
	$\widehat{\boldsymbol{\Sigma}}$	**101**	**103**	**117**	**155**	**156**	**183**	**101**	**103**	**117**	**155**	**156**	**183**
	10	29.7	74.8	134	31.4	79.9	172	27.8	69.9	92.9	29.0	74.4	123
	20	26.6	53.7	95.2	28.9	61.7	136	24.6	49.2	69.3	26.2	55.7	99.3
100	40	24.0	43.6	82.1	27.3	54.3	123	21.7	38.9	64.9	23.7	47.1	95.4
	60	23.8	43.2	89.6	28.3	55.7	133	20.7	37.3	64.6	23.1	45.8	96.1
	80	26.6	49.2	127	33.6	66.1	194	21.1	39.0	66.3	23.6	47.7	102
	av	-	-	-	-	-	-	21.2	38.5	66.7	23.3	46.3	92.4
	$\widehat{\boldsymbol{\Sigma}}$	**51.2**	**52.2**	**59.9**	**78.0**	**77.3**	**94.1**	**51.2**	**52.2**	**59.9**	**78.0**	**77.3**	**94.1**
	25	21.5	35.0	56.4	23.0	41.2	77.9	20.3	31.6	35.4	21.6	37.1	49.1
	50	16.9	24.2	37.8	19.2	31.6	60.2	15.8	21.4	29.3	17.9	28.0	42.4
250	100	14.4	20.7	31.4	17.5	28.4	52.3	13.3	18.1	26.8	15.9	24.5	40.3
	150	14.1	20.8	35.4	17.7	29.0	59.0	12.7	17.8	26.2	15.4	24.1	39.6
	200	15.3	23.4	51.3	20.0	33.6	97.8	12.8	18.8	26.6	15.6	25.2	40.5
	av	-	-	-	-	-	-	13.1	17.9	26.2	15.7	24.2	39.7
	$\widehat{\boldsymbol{\Sigma}}$	**20.6**	**21.0**	**23.8**	**30.9**	**31.7**	**37.6**	**20.6**	**21.0**	**23.8**	**30.9**	**31.7**	**37.6**

[a] Notes: See Notes to Table 1.

16

Table 4: $R_{2,T}(\widetilde{\Sigma})$ for $k = 100$.[a]

		No averaging						Averaging over 20 samples					
		Gaussian			t(8)			Gaussian			t(8)		
		ρ			ρ			ρ			ρ		
n	m	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95
	5	0.38	1.22	4.19	0.41	1.27	4.49	0.32	1.10	3.08	0.34	1.13	3.47
	10	0.38	1.09	2.90	0.42	1.18	3.59	0.32	0.98	1.96	0.33	1.03	2.53
	20	0.38	0.99	2.52	0.44	1.13	3.71	0.31	0.85	1.94	0.33	0.93	2.62
$n = 50$	30	0.41	1.04	2.73	0.51	1.24	3.91	0.31	0.85	2.05	0.33	0.95	2.80
	40	0.52	1.20	3.44	0.68	1.59	4.78	0.33	0.90	2.08	0.35	1.01	2.98
	av	-	-	-	-	-	-	0.30	0.81	1.94	0.32	0.90	2.55
	$\widehat{\Sigma}$	1.99	1.98	2.19	3.11	3.19	3.18	1.99	1.98	2.19	3.11	3.19	3.18
	10	0.34	0.98	2.00	0.36	1.03	2.54	0.30	0.90	1.25	0.31	0.93	1.59
	20	0.32	0.74	1.50	0.34	0.82	2.23	0.28	0.65	1.06	0.29	0.71	1.44
	40	0.31	0.66	1.35	0.35	0.80	2.00	0.26	0.57	1.11	0.28	0.67	1.53
$n = 100$	60	0.33	0.71	1.44	0.39	0.86	2.17	0.27	0.60	1.14	0.29	0.71	1.58
	80	0.39	0.82	1.87	0.48	1.06	2.73	0.28	0.64	1.14	0.31	0.76	1.64
	av	-	-	-	-	-	-	0.26	0.56	1.09	0.28	0.67	1.53
	$\widehat{\Sigma}$	1.00	1.02	1.08	1.51	1.49	1.66	1.00	1.02	1.08	1.51	1.49	1.66
	25	0.26	0.48	0.91	0.27	0.53	1.21	0.23	0.41	0.50	0.25	0.46	0.70
	50	0.21	0.36	0.65	0.24	0.46	0.96	0.19	0.30	0.48	0.21	0.38	0.71
	100	0.21	0.35	0.54	0.24	0.47	0.83	0.18	0.31	0.47	0.21	0.40	0.72
$n = 250$	150	0.22	0.37	0.58	0.27	0.51	0.90	0.19	0.32	0.47	0.22	0.42	0.71
	200	0.25	0.41	0.75	0.32	0.58	1.26	0.20	0.34	0.47	0.24	0.45	0.72
	av	-	-	-	-	-	-	0.18	0.31	0.46	0.21	0.40	0.68
	$\widehat{\Sigma}$	0.41	0.41	0.45	0.61	0.62	0.67	0.41	0.41	0.45	0.61	0.62	0.67

[a] Notes: See Notes to Table 2.

Table 5: $R_{2,S}(\widetilde{\boldsymbol{\Sigma}})$ for $k = 250$.[a]

		No averaging						Averaging over 20 samples					
		Gaussian			t(8)			Gaussian			t(8)		
		ρ			ρ			ρ			ρ		
n	m	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95
100	10	80.8	256	651	84.3	265	734	77.4	249	575	79.5	258	646
	20	78.3	220	432	82.4	233	554	74.5	213	365	76.9	224	462
	40	75.1	184	372	80.7	204	509	70.1	175	309	73.1	190	410
	60	75.5	176	381	82.8	200	537	68.3	162	311	71.3	180	414
	80	81.5	185	474	94.3	223	670	68.6	161	335	72.0	180	450
	av	-	-	-	-	-	-	68.7	169	306	71.6	185	404
	$\widehat{\boldsymbol{\Sigma}}$	**313**	**315**	**334**	**462**	**469**	**513**	**313**	**315**	**334**	**462**	**469**	**513**
250	25	70.5	172	252	73.4	183	333	68.6	167	195	71.0	177	264
	50	62.8	123	188	66.68	140	262	60.7	118	144	63.9	133	208
	100	55.9	98.3	161	61.3	120	235	53.3	93.0	133	57.8	112	192
	150	54.0	95.2	162	61.1	120	243	50.8	88.3	135	55.8	109	195
	200	56.2	100	201	66.0	130	305	50.3	88.5	141	55.6	109	206
	av	-	-	-	-	-	-	52.1	91.8	134	56.7	111	191
	$\widehat{\boldsymbol{\Sigma}}$	**126**	**127**	**133**	**189**	**194**	**207**	**126**	**127**	**133**	**189**	**194**	**207**
500	50	56.7	98.5	128	59.8	111	184	55.3	94.1	91.6	58.2	107	137
	100	45.9	67.1	96.9	50.8	84.8	143	44.5	63.6	73.9	49.0	80.2	110
	200	38.8	56.2	82.4	45.5	74.6	125	37.5	52.7	70.6	43.4	69.5	105
	300	37.5	55.1	83.4	44.9	73.9	128	35.7	50.8	70.8	41.9	67.3	106
	400	38.6	58.0	104	47.6	79.7	166	35.3	51.3	72.4	41.7	67.9	109
	av	-	-	-	-	-	-	36.9	52.1	70.3	42.9	68.8	102
	$\widehat{\boldsymbol{\Sigma}}$	**63.2**	**63.6**	**67.4**	**94.8**	**95.5**	**103**	**63.2**	**63.6**	**67.4**	**94.8**	**95.5**	**103**

18

Table 6: $R_{2,T}(\widetilde{\Sigma})$ for $k = 250$.[a]

		No averaging						Averaging over 20 samples					
		Gaussian			t(8)			Gaussian			t(8)		
		ρ			ρ			ρ			ρ		
n	m	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95
	10	0.35	1.16	3.93	0.37	1.21	4.25	0.32	1.12	3.28	0.33	1.15	3.58
	20	0.35	1.09	2.60	0.37	1.15	3.36	0.32	1.03	2.06	0.33	1.08	2.64
	40	0.35	0.98	2.52	0.38	1.08	3.40	0.31	0.91	2.00	0.33	0.99	2.71
$n = 100$	60	0.37	1.00	2.61	0.41	1.14	3.71	0.31	0.90	2.14	0.33	1.00	2.88
	80	0.42	1.11	3.10	0.50	1.34	4.41	0.32	0.93	2.28	0.34	1.03	3.07
	av	-	-	-	-	-	-	0.30	0.87	1.98	0.32	0.96	2.62
	$\widehat{\Sigma}$	**2.49**	**2.47**	**2.59**	**3.68**	**3.53**	**4.03**	**2.49**	**2.47**	**2.59**	**3.68**	**3.53**	**4.03**
	25	0.31	0.90	1.44	0.32	0.93	1.84	0.30	0.86	1.02	0.30	0.89	1.37
	50	0.29	0.66	1.20	0.31	0.74	1.67	0.27	0.62	0.90	0.28	0.68	1.29
	100	0.28	0.60	1.10	0.30	0.72	1.63	0.26	0.55	0.95	0.27	0.66	1.36
$n = 250$	150	0.29	0.63	1.11	0.32	0.78	1.67	0.26	0.58	0.99	0.28	0.69	1.43
	200	0.32	0.69	1.29	0.37	0.88	2.00	0.27	0.60	1.03	0.29	0.72	1.48
	av	-	-	-	-	-	-	0.25	0.55	0.95	0.27	0.67	1.34
	$\widehat{\Sigma}$	**1.00**	**1.00**	**1.02**	**1.48**	**1.51**	**1.56**	**1.00**	**1.00**	**1.02**	**1.48**	**1.51**	**1.56**
	50	0.26	0.53	0.77	0.27	0.57	1.09	0.25	0.49	0.49	0.26	0.53	0.72
	100	0.22	0.38	0.65	0.24	0.47	0.94	0.21	0.35	0.49	0.23	0.43	0.72
	200	0.21	0.38	0.59	0.24	0.49	0.88	0.20	0.35	0.51	0.23	0.45	0.76
$n = 500$	300	0.22	0.40	0.59	0.26	0.52	0.90	0.21	0.37	0.52	0.24	0.47	0.78
	400	0.24	0.43	0.68	0.29	0.58	1.06	0.22	0.38	0.53	0.25	0.49	0.81
	av	-	-	-	-	-	-	0.20	0.35	0.51	0.23	0.45	0.74
	$\widehat{\Sigma}$	**0.50**	**0.51**	**0.52**	**0.75**	**0.76**	**0.79**	**0.50**	**0.51**	**0.52**	**0.75**	**0.76**	**0.79**

19

Table 7: Average ratio of condition numbers.[a]

		No averaging						Averaging over 20 samples					
		Gaussian			t(8)			Gaussian			t(8)		
		ρ			ρ			ρ			ρ		
n	m	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95	0.5	0.75	0.95
							$k=30$						
50	5	0.036	0.033	0.036	0.027	0.026	0.030	0.019	0.016	0.013	0.013	0.011	0.011
	10	0.043	0.045	0.055	0.033	0.035	0.046	0.023	0.023	0.028	0.016	0.017	0.022
	20	0.057	0.067	0.082	0.043	0.053	0.069	0.030	0.036	0.047	0.020	0.026	0.038
	30	0.079	0.094	0.111	0.061	0.077	0.094	0.037	0.048	0.061	0.026	0.036	0.049
	40	0.155	0.178	0.191	0.123	0.148	0.162	0.050	0.063	0.076	0.035	0.048	0.061
	av	-	-	-	-	-	-	0.029	0.035	0.044	0.020	0.026	0.035
							$k=100$						
250	25	0.068	0.065	0.075	0.047	0.049	0.062	0.048	0.044	0.047	0.031	0.030	0.037
	50	0.081	0.090	0.111	0.057	0.068	0.092	0.058	0.062	0.078	0.037	0.044	0.063
	100	0.104	0.127	0.153	0.074	0.098	0.129	0.074	0.091	0.118	0.049	0.067	0.096
	150	0.130	0.160	0.188	0.094	0.127	0.159	0.087	0.112	0.140	0.058	0.082	0.114
	200	0.178	0.216	0.239	0.132	0.178	0.207	0.103	0.133	0.158	0.069	0.100	0.129
	av	-	-	-	-	-	-	0.073	0.089	0.112	0.047	0.065	0.091
							$k=250$						
500	50	0.035	0.033	0.037	0.021	0.023	0.029	0.028	0.025	0.025	0.016	0.016	0.019
	100	0.042	0.045	0.056	0.025	0.031	0.045	0.032	0.034	0.042	0.018	0.022	0.032
	200	0.053	0.065	0.083	0.033	0.047	0.067	0.040	0.049	0.066	0.023	0.033	0.052
	300	0.065	0.083	0.103	0.041	0.061	0.085	0.048	0.061	0.081	0.028	0.042	0.064
	400	0.086	0.109	0.130	0.056	0.083	0.109	0.055	0.073	0.092	0.033	0.051	0.073
	av	-	-	-	-	-	-	0.041	0.048	0.063	0.023	0.033	0.050

[a] Notes: This Table reports the ratio of condition numbers, $\tilde{c}_{n-m}/\hat{c}_n$, averaged over the number of simulations. Results are for different values of m and ρ, and for different distributions with mean zero and Toeplitz variance matrix $\boldsymbol{\Sigma}$ with typical element given by $\sigma_{ij} = \rho^{|i-j|}$, $i,j = 1,\dots,k$. Shaded cells report the minimum value of the condition number over m. The left panel reports entries for the case where the first m observations in the sample are used to calculate $\hat{\boldsymbol{P}}_1$ and the remaining ones are used to calculate $\tilde{\boldsymbol{\Lambda}}$. The right panel reports entries for the case where we randomly sample the m observations $S = 20$ times and average the resulting estimator to obtain $\tilde{\boldsymbol{\Sigma}}_{m,S}$ as in (12). The line "av" refers to the case where we average the estimator over different values of m, namely $m \in [0.2n, 0.8n]$, and obtain $\tilde{\boldsymbol{\Sigma}}_{M,S}$ as in (13). All results are based on 1,000 simulations.

20

Table 8: Bootstrap-based m_B, for $k = 30, n = 50$, Gaussian distribution.

		ρ	
	0.5	0.75	0.95
Optimal m	20	21	16
Median	24	22	16
Mean	23.5	22.1	16.3
Std. Dev.	1.22	1.61	2.31
Min	20	16	11
10% quantile	22	20	13
25% quantile	23	21	15
75% quantile	24	23	18
95% quantile	25	24	19
Max	27	26	23
Increase in MSE	0.49%	0.05%	0.00

Notes: This table reports results for the bootstrap procedure to choose m in order to minimize the mean squared error (MSE). The first line reports the value of m that minimizes $\mathrm{R}_{2,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}}_m)$. The last line "Increase in MSE" reports the percentage increase in MSE by choosing the bootstrap-based m_B as opposed to the optimal m. For example, for $\rho = 0.5$ the bootstrap suggests taking $m_B = 24$, which results in the MSE being 6.905, while the optimal MSE at $m = 20$ is 6.871, hence an increase of 0.49%. All results are based on 1,000 simulations and 1,000 bootstrap replications.
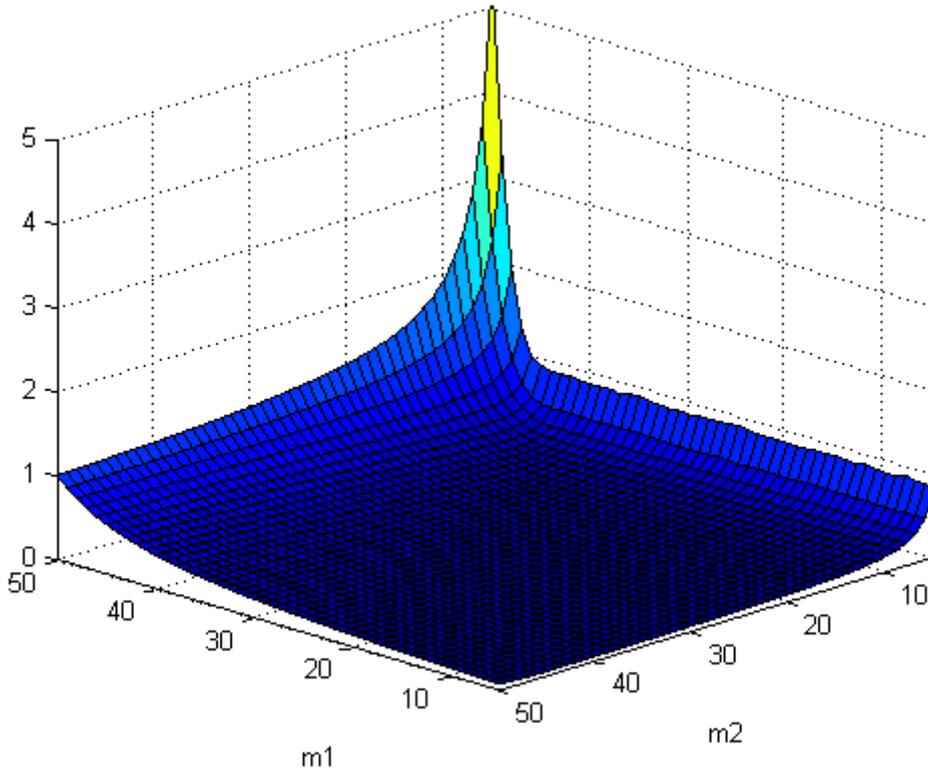
Figure 1: Ratio of condition numbers $\widetilde{c}_{n-m}/\widehat{c}_n$, averaged over simulations, for $k = 30$, $n = 50$, $\rho = 0.5$ and $\boldsymbol{x} \sim \mathrm{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Here, we illustrate the general version of the two-step estimator: we estimate $\boldsymbol{P}_1$ using the first m_1 observations and use the last m_2 (instead of $n - m_1$) observations to calculate $\widetilde{\boldsymbol{\Sigma}}$. In the notation of Proposition 1, $m_2 := n - j$.
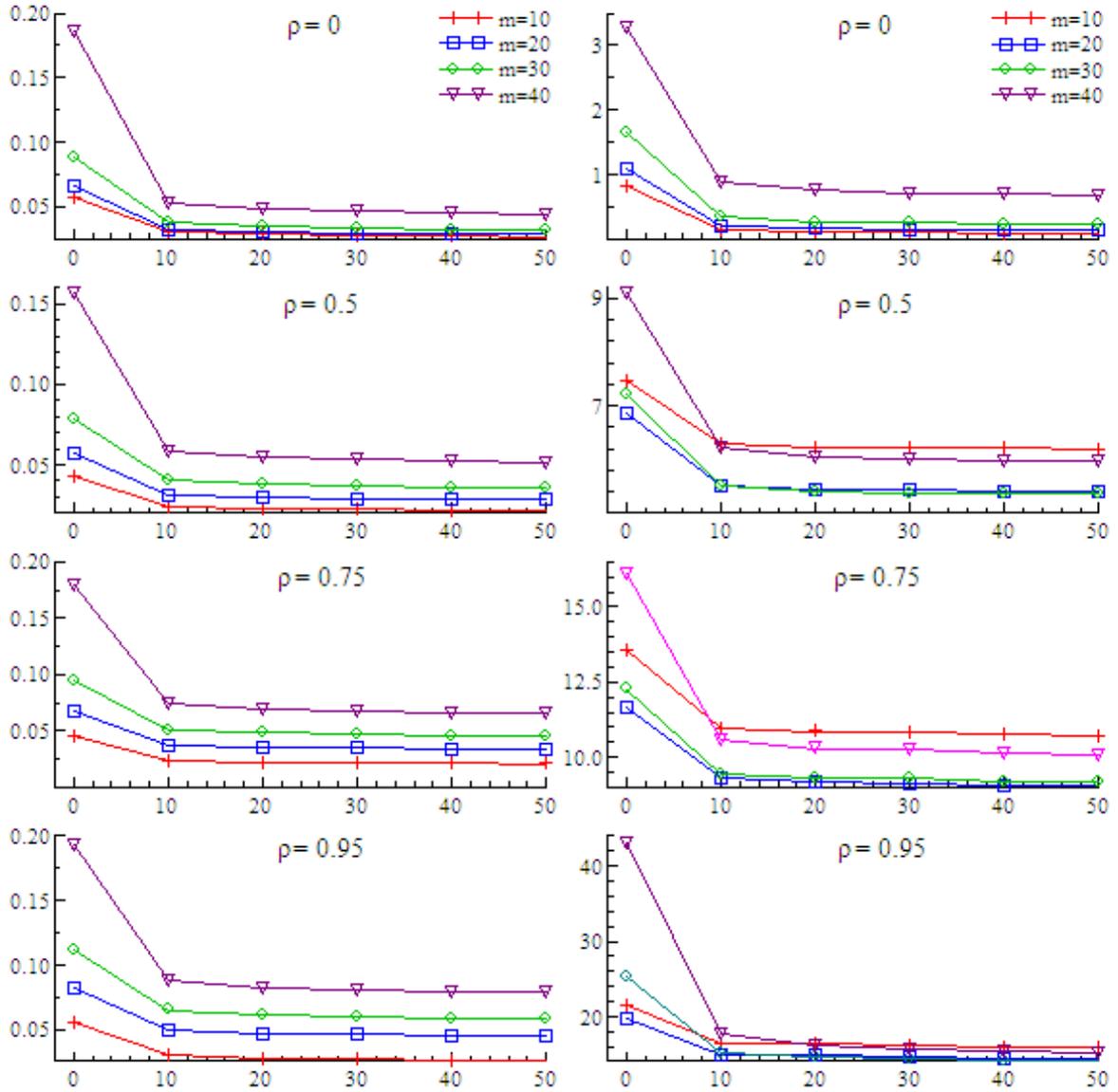
Figure 2: Ratio of condition numbers $\widetilde{c}_{n-m}/\widehat{c}_n$ (left panel) and $\mathrm{R}_{2,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}}_{m,S})$ (right panel), averaged over S simulations (horizontal axes), for $k = 30$, $n = 50$, and $\boldsymbol{x} \sim \mathrm{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma})$.

References

Abadir, K.M. and Magnus, J.R. (2005). *Matrix Algebra.* Cambridge University Press, Cambridge.

Bickel, P.J. and Levina, E. (2008a). Regularized Estimation of Large Covariance Matrices. *Annals of Statistics*, 36, 199–227.

Bickel, P.J. and Levina, E. (2008b). Covariance Regularization by Thresholding. *Annals of Statistics*, 36, 2577–2604.

El Karoui, N. (2009). Operator Norm Consistent Estimation of a Large Dimensional Sparse Covariance Matrices. *Annals of Statistics*, forthcoming.

Engle, R.F., Shephard, N. and Sheppard, K. (2008). Fitting Vast Dimensional Time-Varying Covariance Models. Working paper, University of Oxford.

Fan, J., Fan, Y. and Lv, J. (2008). High Dimensional Covariance Matrix Estimation Using a Factor Model. *Journal of Econometrics*, 147, 186–197.

Fan, J. and Peng, H. (2004). Nonconcave Penalized Likelihood With a Diverging Number of Parameters. *Annals of Statistics*, 32, 928–961.

Huang, J., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance Matrix Selection and Estimation via Penalised Normal Likelihood. *Biometrika*, 93, 85–98.

Jorion, P. (1986). Bayes-Stein Estimation for Portfolio Analysis. *Journal of Financial and Quantitative Analysis*, 21, 279–292.

Lam, C. and Fan, J. (2009). Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation. *Annals of Statistics*, 37, 4254–4278.

Lam, C. and Yao, Q. (2009). Large Precision Matrix Estimation for Time Series Data With Latent Factor Model. Working paper, London School of Economics.

Ledoit, O. and Wolf, M. (2001). Improved Estimation of the Covariance Matrix of Stock Returns With an Application to Portfolio Selection. *Journal of Empirical Finance*, 10, 603–621.

Ledoit, O. and Wolf, M. (2003). A Well-Conditioned Estimator for Large Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, 88, 365–411.

Ledoit, O. and Wolf, M. (2004). Honey, I Shrunk the Sample Covariance Matrix. *Journal of Portfolio Management*, 31, 1–22.

Muirhead, R. (1987). Developments in Eigenvalue Estimation. In Gupta, A.K. (Ed.), *Advances in Multivariate Statistical Analysis*. Reidel, Boston, 277–288.

Rothman, A.J., Bickel, P.J., Levina, E. and Zhu, J. (2008). Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics*, 2, 494–515.

Stein, C. (1956). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In Neyman, J. (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical and Statistical Probability*. University of California, Berkeley, Vol.1, 197–206.

Wang, Y., Li, P., Zou, J. and Yao, Q. (2009). High Dimensional Volatility Modeling and Analysis for High-Frequency Financial Data. Working paper, London School of Economics.

Wu, W.B. and Pourahmadi, M. (2003). Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data. *Biometrika*, 94, 1–17.