Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

DESIGN-FREE ESTIMATION OF LARGE VARIANCE MATRICES*

By Karim M. Abadir, Walter Distaso and Filip Zikes

Imperial College London

This paper introduces a new method for estimating 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 deliver a well-conditioned estimator (by construction), even when there are fewer observations than dimensions. We also show that our estimator has lower error norms than the traditional one. Our estimator is design-free: we make no assumptions on the distribution of the random sample or on any parametric structure the variance matrix may have. Simulations confirm our theoretical results and they also show that our simple estimator does very well in comparison with other existing methods, especially when the data are generated from fat-tailed densities.

1. Introduction.

Apart from calculating the mean, estimating the variance of a random vector is the most basic problem in statistics. It has numerous applications in sciences, social sciences and humanities. 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 the 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 relative to the sample size n. 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)

*This research is supported by the ESRC grants R000239538, RES000230176, RES062230311 and RES062230790.

AMS 2000 subject classifications: Primary 62H12; secondary 62F40

Keywords and phrases: variance matrices, ill-conditioning, resampling, U-statistics

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), Rothman, Bickel, Levina and Zhu (2008), Rothman, Levina and Zhu (2009). 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), where it was applied to the estimation of the mean vector. 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. In order to limit the accumulation error in estimating large variance matrices of asset returns, some recent research has suggested to impose a constraint on a suitably chosen norm of the portfolio weights for asset allocation. These constraints prevent taking extreme positions in single assets, and give rise to well diversified, stable, and sparse portfolios, which tend to perform well out of sample. Examples of this approach are given by Fan, Zhang and Yu (2008), Brodie, De Mol, Daubechies, Giannone and Loris (2009) and DeMiguel, Garlappi, Nogales and Uppal (2009). Interestingly, the resulting weights coming out of the constrained optimization are equivalent to those coming out of the unconstrained optimization after shrinking the variance matrix.

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 (they have the perfect condition number of 1), 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. Effectively, this simple idea reduces the multivariate problem to k univariate ones that have no ill-conditioning difficulties because of orthogonality. Moreover, we improve precision further by repeating our procedure over different subsamples used to estimate the eigenvectors, and we show that averaging these leads to a superior estimator.

Even though we only use the simple traditional formula for the sample variance matrix in both steps of our basic procedure, the result is a well-conditioned and precise estimator. 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: $k > n$. Our estimator outperforms the traditional one, not only by achieving a substantial improvement in the condition number, but also by large improvements in error norms that measure its deviation from the true variance matrix. 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). We also show that our simple estimator does very well in comparison to other existing methods, especially when the data are generated from fat-tailed densities.

Our method has a number of other attractive features. First, it is design-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. Such operations are trivially easy to implement in our setup, since matrix functions are efficiently written in terms of eigenvalues and eigenvectors; e.g. see Abadir and Magnus (2005, Ch. 9).

This paper is organized as follows. Section 2 introduces the proposed estimator in its simple then general versions, and establishes its main properties. Section 3 studies in a Monte-Carlo experiment the finite-sample properties of our estimator and how it compares with other methods. It also 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. Second, we investigate the estimator's properties and describe the optimal choice of two subsampling parameters. We do so for the simplest baseline formulation of our estimator, the full version of it being studied at the end of the section.

2.1. *The setup and the idea behind the estimator.* Let $\boldsymbol{\Sigma} := \operatorname{var}(\boldsymbol{x})$ 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 relative to n)

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

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

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{P}}', \tag{2.1}$$

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 $\widehat{\boldsymbol{\Sigma}}$. 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 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

$$\widehat{\boldsymbol{\Lambda}} = \widehat{\boldsymbol{P}}'\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{P}} = \operatorname{diag}(\widehat{\operatorname{var}}(\widehat{\boldsymbol{p}}_1'\boldsymbol{x}), \ldots, \widehat{\operatorname{var}}(\widehat{\boldsymbol{p}}_k'\boldsymbol{x})) \tag{2.2}$$

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 $\widehat{\boldsymbol{p}}_i'\boldsymbol{x}$ in (2.2)) 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. Propositions 2–3 will show that the precision of the new estimator is optimized by expressing m as a function of n asymptotically, and this will be followed by a discussion of how to calculate the optimal m by resampling in any finite sample.

Intuitively, by orthogonalizing the data, our estimator reduces the multivariate problem of ill-conditioning and imprecision to a univariate one for each of the diagonal elements of (2.2), for which there is a simple positive definite solution even by traditional methods of estimation. The result is a well-conditioned estimator of $\boldsymbol{\Sigma}$, even when $k \geq n$ and the traditional $\widehat{\boldsymbol{\Sigma}}$ is not positive definite.

Another advantage of our procedure is that we can estimate the matrix itself as well as its inverse or any function thereof in one go from the eigenvalue decomposition. The other methods seen in the introduction focus on the variance matrix, and if inverse is needed, one has to make further elaborate calculations to obtain it. This is computationally costly and imprecise if the dimension is large. In addition to the advantages seen so far, we will show that also the precision of our estimator is an advantage, even though we only use the traditional sample variance estimator in both steps of our procedure.

2.2. *Investigation of the baseline estimator's properties and its general version.* To summarize the procedure in equations, we start by writing

$$\boldsymbol{X}' = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) =: (\boldsymbol{X}_1', \boldsymbol{X}_2')\,, \tag{2.3}$$

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

$$\widehat{\boldsymbol{\Sigma}}_1 := \frac{1}{m}\boldsymbol{X}_1'\boldsymbol{M}_m\boldsymbol{X}_1 = \widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{\Lambda}}_1\widehat{\boldsymbol{P}}_1', \tag{2.4}$$

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

$$\operatorname{dg}\left(\widehat{\operatorname{var}}(\widehat{\boldsymbol{P}}_1'\boldsymbol{x})\right) \equiv \operatorname{dg}\left(\widehat{\boldsymbol{P}}_1'\widehat{\boldsymbol{\Sigma}}_2\widehat{\boldsymbol{P}}_1\right) = \frac{1}{n-m}\operatorname{dg}\left(\widehat{\boldsymbol{P}}_1'\boldsymbol{X}_2'\boldsymbol{M}_{n-m}\boldsymbol{X}_2\widehat{\boldsymbol{P}}_1\right) =: \widetilde{\boldsymbol{\Lambda}} \tag{2.5}$$

to replace $\widehat{\boldsymbol{\Lambda}}$ of (2.1) and obtain the new estimator

$$\widetilde{\boldsymbol{\Sigma}} := \widehat{\boldsymbol{P}}\widetilde{\boldsymbol{\Lambda}}\widehat{\boldsymbol{P}}' = \widehat{\boldsymbol{P}}\,\mathrm{dg}\left(\widehat{\boldsymbol{P}}_1'\widehat{\boldsymbol{\Sigma}}_2\widehat{\boldsymbol{P}}_1\right)\widehat{\boldsymbol{P}}'. \tag{2.6}$$

Note that we use the simple traditional estimator of variance matrices $\widehat{\mathrm{var}}(\cdot)$ in each of the two steps of our procedure. 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 $\mathrm{dg}(\widehat{\mathrm{var}}(\widehat{\boldsymbol{P}}_1'\boldsymbol{x}))$ rather than $\widehat{\mathrm{var}}(\widehat{\boldsymbol{P}}_1'\boldsymbol{x})$ in the second step given by (2.5)–(2.6), since by doing so we impose the correct restriction that estimators of $\boldsymbol{\Lambda}$ should be diagonal, restricting off-diagonal elements to be zero. Second, the estimate $\widetilde{\boldsymbol{\Sigma}}$ is almost surely nonsingular, like the true $\boldsymbol{\Sigma}$, because of the use of dg in (2.5). Third, we choose to demean $\boldsymbol{X}_2$ by its own mean (rather than the whole sample's mean) mainly for robustness considerations in practice, 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 (2.5). 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 2.1. *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* $\boldsymbol{x}$ *exists and that* $\boldsymbol{\Sigma}$ *is positive definite. 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 (2.5) 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 we conjecture that it is not a necessary condition.

Proof. For $m > j + 2$,

$$\widetilde{\boldsymbol{\Lambda}}_j = \frac{1}{n-j}\,\mathrm{dg}\left(\widehat{\boldsymbol{P}}_1' \sum_{i=j+1}^{m} \boldsymbol{y}_i\boldsymbol{y}_i'\widehat{\boldsymbol{P}}_1\right) + \frac{1}{n-j}\,\mathrm{dg}\left(\widehat{\boldsymbol{P}}_1' \sum_{i=m+1}^{n} \boldsymbol{y}_i\boldsymbol{y}_i'\widehat{\boldsymbol{P}}_1\right) \tag{2.7}$$

$$= \frac{m-j}{n-j}\,\mathrm{dg}\,(\boldsymbol{S}_j) + \frac{n-m}{n-j}\widetilde{\boldsymbol{\Lambda}}_m,$$

which is a weighted average of $\mathrm{dg}\,(\boldsymbol{S}_j)$ 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 (2.4), 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). See Anderson (1963) and Davis (1977) for the asymptotic normality of the traditional estimator on which our procedure is based.

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 (2.7), the two limiting matrices for the components in (2.7) 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 propositions.

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

$$c_n = \frac{\lambda_1 + \omega_1\delta_n}{\lambda_k - \omega_k\delta_n} = \left(c + \frac{\omega_1\delta_n}{\lambda_k}\right)\left(1 + \frac{\omega_k\delta_n}{\lambda_k} + O(\delta_n^2)\right) \tag{2.8}$$

$$= c + \frac{\omega_1 + c\omega_k}{\lambda_k}\delta_n + O(\delta_n^2).$$

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 (2.7) and dropping the remainder terms, we get the asymptotic condition number of $\widetilde{\boldsymbol{\Lambda}}_j$ as

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

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 $m \to \infty$. □

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 start by assuming that k is fixed, but we relax this condition by the end of this section. However, we will need to assume the existence of fourth-order moments for $\boldsymbol{x}$ when we consider l_2-norm precision criteria for the estimation of $\boldsymbol{\Sigma}$.

Define the following criteria that are inversely related to the precision of the new estimator $\widetilde{\boldsymbol{\Sigma}}$:

$$\text{(2.9)} \qquad \mathrm{R}_l(\widetilde{\boldsymbol{\Sigma}}) := \mathrm{E}(||\operatorname{vec}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})||_l^l), \qquad l = 1, 2,$$

and

$$\text{(2.10)} \qquad \mathrm{R}_{l,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}}) := \mathrm{E}(||\operatorname{vech}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})||_l^l), \qquad l = 1, 2,$$

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 $l = 1$ and $l = 2$, respectively. The half-vec operator, vech, selects only the distinct elements of a general symmetric matrix.

Since so far the dimension k is finite as n increases, there is asymptotically no difference in considering the usual or the S 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 the estimators of $\boldsymbol{\Sigma}$. Also, we will analyze the effect of varying k on these norms after the following two propositions.

Proposition 2.2. *As $n - m \to \infty$ and $m \to \infty$, the precision criteria in (2.9)–(2.10) are optimized asymptotically for $\widetilde{\boldsymbol{\Sigma}}$ by taking $m/\sqrt{n} \to \infty$ and $m/n \to 0$, if the positive definite $\boldsymbol{\Sigma}$ is not a scalar matrix.*

Proof. Under the i.i.d. assumption and the existence of fourth moments, $\widehat{\boldsymbol{\Sigma}}_2$ satisfies the CLT

$$\widehat{\boldsymbol{\Sigma}}_2 = \boldsymbol{\Sigma} + \frac{1}{\sqrt{n-m}} \boldsymbol{Z}_2 \left(1 + o_p\left(1\right)\right),$$

where the elements of $\boldsymbol{Z}_2$ are jointly normal with mean zero and some finite positive definite variance matrix; see Anderson (1963) for the case of $\boldsymbol{x}$ normal and Davis (1977) for its generalization. Define

$$\widehat{\boldsymbol{\Omega}}_1 := \boldsymbol{P}' \widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{P},$$

whose eigenvalues are the same as those of $\widehat{\boldsymbol{\Sigma}}_1$ and its eigenvectors are $\widehat{\boldsymbol{Q}}_1 := \boldsymbol{P}' \widehat{\boldsymbol{P}}_1$. We can write $\widehat{\boldsymbol{\Omega}}_1 = \widehat{\boldsymbol{Q}}_1 \widehat{\boldsymbol{\Lambda}}_1 \widehat{\boldsymbol{Q}}_1'$ which satisfies the CLT

$$\widehat{\boldsymbol{\Omega}}_1 = \boldsymbol{\Lambda} + \frac{1}{\sqrt{m}} \boldsymbol{U}_1 \left(1 + o_p\left(1\right)\right),$$

where the elements of $\boldsymbol{U}_1$ are jointly normal with mean zero and some positive definite variance matrix (these elements are uncorrelated when $\boldsymbol{x}$ is normal).

Let $\boldsymbol{\Sigma}$ have r distinct eigenvalues, $\lambda_1, \lambda_2, \dots, \lambda_r$, all positive, with multiplicities $k_1, k_2, \dots, k_r$. The two extreme cases are all eigenvalues identical ($r = 1$ and $k_1 = k$) and all eigenvalues distinct ($r = k$ and $k_i = 1$, $i = 1, \dots, k$). Now partition the matrices $\boldsymbol{U}_1$ and $\widehat{\boldsymbol{Q}}_1$ into submatrices with $k_1, \dots, k_r$ rows and columns as $\boldsymbol{U}_1 = (\boldsymbol{U}_{1,ij})$ and $\widehat{\boldsymbol{Q}}_1 = (\widehat{\boldsymbol{Q}}_{1,ij})$, where $i, j = 1, \dots, r$. Also, partition $\widehat{\boldsymbol{P}}_1 = (\widehat{\boldsymbol{P}}_{1,i})$ and $\boldsymbol{P} = (\boldsymbol{P}_i)$ as blocks of $k_1, \dots, k_r$ columns. Then, by defining the normalized $\widehat{\boldsymbol{R}}_{1,ij} := m^{1/2} \widehat{\boldsymbol{Q}}_{1,ij}$ (for $i \neq j$) like Anderson (1963), we have

$$\widehat{\boldsymbol{P}}_{1,i} = \boldsymbol{P}_i \widehat{\boldsymbol{Q}}_{1,ii} + m^{-1/2} \sum_{j \neq i} \boldsymbol{P}_j \widehat{\boldsymbol{R}}_{1,ji}.$$

Turning to our estimator of $\boldsymbol{\Lambda}$, partition it into the block diagonal $\widetilde{\boldsymbol{\Lambda}} = \operatorname{diag}(\dots, \widetilde{\boldsymbol{\Lambda}}_i, \dots)$, where $\widetilde{\boldsymbol{\Lambda}}_i$ is the $k_i \times k_i$ matrix corresponding to the eigenvalue λ_i with multiplicity k_i. Then, by direct substitution, we get the leading terms of the asymptotic expansion

$$
\begin{aligned}
\widetilde{\boldsymbol{\Lambda}}_i &= \mathrm{dg}(\widehat{\boldsymbol{P}}'_{1,i}\widehat{\boldsymbol{\Sigma}}_2\widehat{\boldsymbol{P}}_{1,i}) \\
&\overset{a}{=} \mathrm{dg}((\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} + m^{-1/2}\sum_{j\neq i}\boldsymbol{P}_j\widehat{\boldsymbol{R}}_{1,ji})'(\boldsymbol{\Sigma} + (n-m)^{-1/2}\boldsymbol{Z}_2) \\
&\qquad\qquad\qquad\qquad \times (\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} + m^{-1/2}\sum_{j\neq i}\boldsymbol{P}_j\widehat{\boldsymbol{R}}_{1,ji})) \\
&= \mathrm{dg}(\widehat{\boldsymbol{Q}}'_{1,ii}\boldsymbol{P}'_i\boldsymbol{\Sigma}\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} + (n-m)^{-1/2}\widehat{\boldsymbol{Q}}'_{1,ii}\boldsymbol{P}'_i\boldsymbol{Z}_2\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii}
\end{aligned}
$$

$$
+m^{-1/2}\sum_{j\neq i}\widehat{\boldsymbol{R}}'_{1,ji}\boldsymbol{P}'_j\boldsymbol{\Sigma}\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} + m^{-1/2}(n-m)^{-1/2}\sum_{j\neq i}\widehat{\boldsymbol{R}}'_{1,ji}\boldsymbol{P}'_j\boldsymbol{Z}_2\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} \tag{2.11}
$$

$$
+m^{-1/2}\sum_{j\neq i}\widehat{\boldsymbol{Q}}'_{1,ii}\boldsymbol{P}'_i\boldsymbol{\Sigma}\boldsymbol{P}_j\widehat{\boldsymbol{R}}_{1,ji} + m^{-1/2}(n-m)^{-1/2}\sum_{j\neq i}\widehat{\boldsymbol{Q}}'_{1,ii}\boldsymbol{P}'_i\boldsymbol{Z}_2\boldsymbol{P}_j\widehat{\boldsymbol{R}}_{1,ji} \tag{2.12}
$$

$$
+m^{-1}\sum_{j\neq i}\sum_{l\neq i}\widehat{\boldsymbol{R}}'_{1,ji}\boldsymbol{P}'_j\boldsymbol{\Sigma}\boldsymbol{P}_l\widehat{\boldsymbol{R}}_{1,li} + m^{-1}(n-m)^{-1/2}\sum_{j\neq i}\sum_{l\neq i}\widehat{\boldsymbol{R}}'_{1,ji}\boldsymbol{P}'_j\boldsymbol{Z}_2\boldsymbol{P}_l\widehat{\boldsymbol{R}}_{1,li}). \tag{2.13}
$$

Using $\boldsymbol{P}'\boldsymbol{\Sigma}\boldsymbol{P} = \boldsymbol{\Lambda}$, we have $\widehat{\boldsymbol{Q}}'_{1,ii}\boldsymbol{P}'_i\boldsymbol{\Sigma}\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} = \lambda_i\widehat{\boldsymbol{Q}}'_{1,ii}\widehat{\boldsymbol{Q}}_{1,ii} = \lambda_i\boldsymbol{I}_{k_i} - m^{-1}\lambda_i\widehat{\boldsymbol{W}}_{1,ii}$ where $\widehat{\boldsymbol{W}}_{1,ii}$ is a sum obtainable as in Anderson (1963, p.128). The off-diagonals of $\boldsymbol{P}'\boldsymbol{\Sigma}\boldsymbol{P}$ being zero, the first term in each of (2.11) and (2.12) is zero, and the double sums in (2.13) can be written as single sums. The second terms in (2.11), (2.12) and (2.13) will always be of smaller order than the remaining terms, so we can drop them. We are therefore left with

$$
\begin{aligned}
\widetilde{\boldsymbol{\Lambda}}_i - \lambda_i\boldsymbol{I}_{k_i} \overset{a}{=}\ & (n-m)^{-1/2}\,\mathrm{dg}(\widehat{\boldsymbol{Q}}'_{1,ii}\boldsymbol{P}'_i\boldsymbol{Z}_2\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii}) \\
& + m^{-1}\sum_{j\neq i}\lambda_j\,\mathrm{dg}(\widehat{\boldsymbol{R}}'_{1,ji}\widehat{\boldsymbol{R}}_{1,ji}) - m^{-1}\lambda_i\,\mathrm{dg}(\widehat{\boldsymbol{W}}_{1,ii}),
\end{aligned} \tag{2.14}
$$

where $\lambda_{\cdot} \neq 0$ and, in general, the last two terms don't cancel and the diagonal terms are not all zero. From this we can deduce the following four cases:
(a) If $m/\sqrt{n} \to \infty$ and $m/n \to \gamma$, where $0 \leq \gamma < 1$, the leading term of the expansion is the first one.
(b) If $m/\sqrt{n} \to \infty$ and $m/n \to 1$ (with $n - m \to \infty$), the leading term of the expansion is the first one but its order is larger than $n^{-1/2}$ hence suboptimal compared to case (a).
(c) If $m/\sqrt{n} \to \gamma$, where $0 < \gamma < \infty$, all terms are of the same order and we have the same convergence rate as in case (a). Note that the estimator will have a finite-sample bias of order $n^{-1/2}$ in this case.

(d) If $m/\sqrt{n} \to 0$, the leading term of the expansion is not the first one, but its order is larger than $n^{-1/2}$ hence suboptimal.

Since $\widehat{\boldsymbol{Q}}_{1,ii}\widehat{\boldsymbol{Q}}'_{1,ii} \xrightarrow{p} \boldsymbol{I}_{k_i}$, where the limit is independent of m, the optimal case for $\widetilde{\boldsymbol{\Lambda}}$ is (a) with $\gamma = 0$. Since $\widehat{\boldsymbol{P}}$ is based on the full sample and hence does not depend on m, we get the same optimal m for $\widetilde{\boldsymbol{\Sigma}}$ as for $\widetilde{\boldsymbol{\Lambda}}$. □

The case of a scalar matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_k$ (which is rare) is essentially the case of scalar estimation of one variance σ^2. In (2.15), the sums included in the $O_p(m^{-1})$ terms are empty sums when $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_k = \lambda \boldsymbol{I}_k$, so the optimal is to take $n - m$ as large as possible to minimize the leading term. The result is obtained more easily and more generally for finite samples as follows. We have $\boldsymbol{Q}(\sigma^2 \boldsymbol{I}_k)\boldsymbol{Q}' = \sigma^2 \boldsymbol{I}_k$ for *any* orthogonal $\boldsymbol{Q}$. In other words, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_k$ implies that the precision of its estimation is invariant to $\boldsymbol{P}$ and the optimal choice of m is as small as possible ($m = 2$) to increase the precision of the eigenvalues estimated in the second step. We will not consider scalar matrices henceforth.

Up to now, we constructed our simple 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 $q := \binom{n}{m}$ possibilities of choosing the m observations to calculate $\widehat{\boldsymbol{P}}_1$, all of them being equivalent due to the i.i.d. assumption. Averaging our estimates of $\widetilde{\boldsymbol{\Lambda}}$ over these different possibilities implies that we can write the leading term of (2.15) as a U-statistic (that is centred around zero by the independence of the subsamples generating $\boldsymbol{Z}_2$ and $\widehat{\boldsymbol{Q}}_{1,ii}$) and reduce its magnitude as we will show. The sampling intuition behind this is that averaging will reduce the variability that comes with the choice of any one specific combination of m observations.

To define our general estimator, let $\boldsymbol{X}'_s := (\boldsymbol{X}'_{1,s}, \boldsymbol{X}'_{2,s})$, where $\boldsymbol{X}'_{1,s} := \left(\boldsymbol{x}^s_{1,1}, \ldots, \boldsymbol{x}^s_{1,m}\right)$ is obtained by randomly sampling without replacement m columns from the original $\boldsymbol{X}'$, and $\boldsymbol{X}'_{2,s} := \left(\boldsymbol{x}^s_{2,1}, \ldots, \boldsymbol{x}^s_{2,n-m}\right)$ 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

$$\widetilde{\boldsymbol{\Sigma}}_{m,S} := \frac{1}{S}\sum_{s=1}^{S} \widetilde{\boldsymbol{\Sigma}}_{m,s} \tag{2.15}$$

denote the average over S samples. Computational burden makes the choice $S = q$ prohibitive for large n, but we will see in the next section that a relatively small number of samples S suffices to reap most of the benefits of averaging. Except for Proposition 2, the properties derived earlier for our simpler estimator apply also to our general estimator (2.15). As for the choice of m, we get the following result to complement Proposition 2.

Proposition 2.3. *As $n-m\to\infty$ and $m\to\infty$, the precision criteria in (2.9)–(2.10) are optimized asymptotically for the general estimator $\widetilde{\boldsymbol{\Sigma}}_{m,q}$ by taking $m/n\to\gamma$ with $\gamma\in(0,1)$, if the positive definite $\boldsymbol{\Sigma}$ is not a scalar matrix.*

Proof. Following Anderson (1963) and Davis (1977), we have $\widehat{\boldsymbol{Q}}_{1,ii}=\boldsymbol{Q}_{ii}+O_p(m^{-1/2})$ where $\boldsymbol{Q}_{ii}$ is an orthogonal matrix having the conditional Haar invariant distribution independently of $\boldsymbol{Z}_2$, as they are based on independent samples. This simplifies the asymptotic expansion in (2.15) to

$$\widetilde{\boldsymbol{\Lambda}}_i-\lambda_i\boldsymbol{I}_{k_i}\overset{a}{=}(n-m)^{-1/2}\operatorname{dg}(\boldsymbol{Q}'_{ii}\boldsymbol{P}'_i\boldsymbol{Z}_2\boldsymbol{P}_i\boldsymbol{Q}_{ii})+((n-m)m)^{-1/2}\boldsymbol{\Delta}\\+m^{-1}\sum_{j\neq i}\lambda_j\operatorname{dg}(\widehat{\boldsymbol{R}}'_{1,ji}\widehat{\boldsymbol{R}}_{1,ji})-m^{-1}\lambda_i\operatorname{dg}(\widehat{\boldsymbol{W}}_{1,ii}),\tag{2.16}$$

where $\boldsymbol{\Delta}=O_p(1)$. The independence of $\boldsymbol{Q}_{ii}$ and $\boldsymbol{Z}_2$ allows an asymptotic U-statistic representation of the first term of the expansion when resampling and averaging. Let us start by conditioning on $\boldsymbol{Q}_{ii}$. Since $\boldsymbol{Z}_2$ arises from the CLT for a sum of $n-m$ i.i.d. observations, we have the U-statistic kernel for its elements

$$h:=(n-m)^{-1/2}\left(s_{i_1}+\cdots+s_{i_{n-m}}\right)$$

with $i_1,\dots,i_{n-m}$ integers taken from $\{1,\dots,n\}$, hence $(n-m)^{-1}\operatorname{var}(s_1+\cdots+s_l+c)\propto l/(n-m)$ with $c:=(n-m)^{-1/2}\operatorname{E}(s_{l+1}+\cdots+s_{n-m})$. Then, the variance of our U-statistic is proportional to

$$\binom{n}{m}^{-1}\sum_{l=1}^{n-m}\binom{n-m}{l}\binom{m}{n-m-l}\frac{l}{n-m}=\frac{n-m}{n}\tag{2.17}$$

and the leading term of the expansion of the averaging estimator has to be normalized by the root of this fraction, so the term becomes $O_p(n^{-1/2})$ independently of m, when we condition on $\boldsymbol{Q}_{ii}$. This variance-reduction factor is also unconditional (by $\operatorname{var}(y)=\operatorname{var}_q(\operatorname{E}_{z|q}y)+\operatorname{E}_q\operatorname{var}_{z|q}(y)=\operatorname{E}_q\operatorname{var}_z(y)$) because $\boldsymbol{Z}_2$ is centred around zero and $\boldsymbol{Q}_{ii}$ is distributed uniformly on the unit sphere.

The leading term is independent of m as a result of resampling and averaging, so the choice of optimal m will be decided by equalizing the order of magnitude of the next terms: if one is larger than the others, then it can be reduced until equality is achieved. The third and fourth terms of (2.16) have positive-definite matrices, so resampling and averaging has no effect on the order of magnitude which is still $O_p(m^{-1})$. Unlike the first term, the second one is not a U-statistic: $\boldsymbol{\Delta}$ contains $(\widehat{\boldsymbol{Q}}_{1,ii} - \boldsymbol{Q}_{ii})$ and $\boldsymbol{Z}_2$ that are independent for any given split of the sample, but the next resample will lead to a $\boldsymbol{Z}_2$ that is correlated with the previous $\widehat{\boldsymbol{Q}}_{1,ii}$ and so on. As a result, all the $\binom{n}{m}$ terms will be correlated and there is no fraction of reduction as in (2.17). Equalizing the orders $((n-m)m)^{-1/2}$ and m^{-1} leads to $m/n \to \gamma$ with $\gamma \in (0,1)$ as optimal for averaging the $\widetilde{\boldsymbol{\Lambda}}$'s over $s = 1, \dots, q$. Since $\widehat{\boldsymbol{P}}$ is based on the full sample and hence does not depend on m, averaging over $\widetilde{\boldsymbol{\Lambda}}_{m,s}$ or $\widetilde{\boldsymbol{\Sigma}}_{m,s}$ as in (2.15) leads to the same optimal m. □

The leading term is now $O_p(n^{-1/2})$ independently of m, which we will show in the next section to give a very stable performance of our general estimator as we vary m/n within a wide range of asymptotic proportionality factors γ. Furthermore, the leading terms are now smaller than in the case of our simple estimator $\widetilde{\boldsymbol{\Sigma}}$. When we use $\widetilde{\boldsymbol{\Sigma}}$ and do not average, applying $\widehat{\boldsymbol{Q}}_{1,ii} = \boldsymbol{Q}_{ii} + O_p(m^{-1/2})$ to the asymptotic expansion (2.15) gives

$$\widetilde{\boldsymbol{\Lambda}}_i - \lambda_i \boldsymbol{I}_{k_i} = O_p\left(\frac{1}{\sqrt{n-m}}\right) + O_p\left(\frac{1}{\sqrt{(n-m)\,m}}\right) + O_p\left(\frac{1}{m}\right)$$

and by choosing $m/n \to 0$ with m as large as possible as in Proposition 2, this becomes

$$\widetilde{\boldsymbol{\Lambda}}_i - \lambda_i \boldsymbol{I}_{k_i} = O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{\sqrt{nm}}\right) + O_p\left(\frac{1}{m}\right) = O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{m}\right),$$

where the last $O_p(m^{-1})$ term is bigger order than with averaging because the optimal m is such that $m/n \to 0$ in Proposition 2 while $m/n \to \gamma \in (0,1)$ in Proposition 3.

Why is there an increase in the precision of the averaging estimator $\widetilde{\boldsymbol{\Sigma}}_{m,q}$ compared to the traditional estimator $\widehat{\boldsymbol{\Sigma}}$? Without repeating the algebra of the proof of Proposition 2, we can write down the expansion for the traditional estimator as

$$\widehat{\boldsymbol{\Lambda}}_i - \lambda_i \boldsymbol{I}_{k_i} \overset{a}{=} n^{-1/2} \widehat{\boldsymbol{Q}}_{ii}' \boldsymbol{P}_i' \boldsymbol{Z} \boldsymbol{P}_i \widehat{\boldsymbol{Q}}_{ii} + n^{-1} \sum_{j \neq i} \lambda_j \widehat{\boldsymbol{R}}_{ji}' \widehat{\boldsymbol{R}}_{ji} - n^{-1} \lambda_i \widehat{\boldsymbol{W}}_{ii},$$

where $\boldsymbol{Z}$ is normal like $\boldsymbol{Z}_2$ was. No resampling-and-averaging will reduce the leading term here, because $\boldsymbol{Z}$ and $\widehat{\boldsymbol{Q}}_{ii}$ are (cor)related in general, unlike in the case of our procedure's decomposition into eigenvectors and eigenvalues from independent subsamples.[1] Furthermore, this correlation means that the leading term is not centred around zero when n is finite.

Regarding the consistency of our estimators as we vary k as well as n, our simple estimator has a k/m term with $m/n \to 0$ and m as large as possible, while our averaging estimator has a k/n term. Using the element-wise definition of convergence, consistency requires $k = o(n)$ for both estimators. If we require convergence in mean square ($l = 2$ in our notation), then the orders under our norms in (2.9)–(2.10) would be $k = o(\sqrt{m})$ for the simple estimator, and $k = o(\sqrt{n})$ for the averaging one, where the latter allows k to grow faster than the former. If we take $l = 1$, they become $k = o(\min\{n^{1/4}, m^{1/3}\}) = o(n^{1/4})$ and $k = o(\min\{n^{1/4}, n^{1/3}\}) = o(n^{1/4})$, respectively. This is rather slow and it is a stringent criterion because it deals with sums of absolute values rather than the squares of the elements.

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, not just asymptotics. 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 estimator and to compare its performance with its most popular competitors. We are also 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). In Subsection 3.1, we employ the symmetric-matrix precision criteria defined in (2.10) in our comparisons. In Subsection 3.2, we investigate the reduction in the condition number of the two-step estimator ($\widetilde{c}_{n-m}$) relative to the sample variance matrix benchmark ($\widehat{c}_n$). Finally, Subsection 3.3 investigates the automation of the choice of m.

Our simulation design is as follows. We let the true variance matrix $\boldsymbol{\Sigma}$ have a Toeplitz structure with typical element given by $\sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \dots, k$, where $k \in \{30, 100, 250\}$ 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 $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_k$; see the discussion following Proposition 2. The random vector $\boldsymbol{x}$ is drawn either from the normal distribution or from the multivariate Student t distribution with five degrees of freedom, denoted t(5), with population

[1] Sampling with replacement would be inferior to $\widetilde{\boldsymbol{\Sigma}}_{m,q}$ (which samples without replacement) for the same reason of overlap in $\boldsymbol{Z}_{\cdot}$ and $\widehat{\boldsymbol{Q}}_{\cdot ii}$, and because of the increase in condition numbers that arises from reusing observations (see Proposition 1).

mean equal to zero without loss of generality. All simulations are based on 1,000 Monte Carlo replications, to save computational time. For example, we repeated the calculations for $k = 30$ with 10,000 replications and essentially identical results were obtained.

Recalling the definition of our general estimator (2.15), we illustrate the choice of S with a preliminary simulation. In Figure 1, we vary S on the horizontal axis (with the simple no-averaging baseline case of our estimator at the origin of the axis) and report the corresponding changes in our estimator's condition number $\widetilde{c}_{n-m}$ relative to the traditional estimator's $\widehat{c}_n$, as well as changes to the precision $\mathrm{R}_{2,\mathrm{S}}$ of $\widetilde{\boldsymbol{\Sigma}}_{m,S}$. We can see that, whatever the choice of m which we will analyze later, the benefits to be achieved occur very quickly for small S and there is not much to be gained from taking a large S, so we use $S = 20$ henceforth. This is true for various values of ρ, and we simulated $S = 10, 20, \ldots, 100$ but only plotted up to 50. The same pattern of results also repeats for different n, k, and distributions.

3.1. *Estimator's Precision.* The results are summarized in Table 1, where the line labelled "av" will be analyzed two paragraphs below. For now we focus on the lines for the traditional estimator $\widehat{\boldsymbol{\Sigma}}$ and our $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ (the rest of the table). The shaded boxes highlight the best-performing case. We see that our two-step estimator is always more precise except in one case: when ρ equals the extreme 0.95 *and* the data are Gaussian. Otherwise, the achieved reduction in the mean squared error is very large, although decreasing with ρ, and is more pronounced for data generated from the fat-tailed Student t distribution.[2] The gains are truly massive, e.g. $\mathrm{R}_{2,\mathrm{S}}$ is better for our estimator by a factor of *eleven* in the case of t(5) and $\rho = 0.5$ for $n = 100$ and $k = 250$.

Throughout the tables, we see a robust performance as m varies around its optimal value, more specifically around approximately $m \in [0.2n, 0.8n]$; recall the asymptotic proportionality of m to n, which was obtained in Proposition 3 and was discussed immediately afterwards. This suggests to construct an estimator based on averaging $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ over $m \in [0.2n, 0.8n]$. This "grand average" estimator is defined as

$$\widetilde{\boldsymbol{\Sigma}}_{M,S} := \frac{1}{M} \sum_{m \in \mathcal{M}} \widetilde{\boldsymbol{\Sigma}}_{m,S}, \tag{3.1}$$

where M is the number of elements in the grid $\mathcal{M} := \{m_1, m_2, \ldots, m_M\}$, where $1 \leq m_1 \leq \cdots \leq m_M \leq n$. Results are reported in the line labelled

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

"av". 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.

For the alternative precision measures, R_2, R_1, and $\mathrm{R}_{1,\mathrm{S}}$, 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(5)), 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.

We also compare the performance of our estimator with its most popular competitors in Table 2. We consider the shrinkage towards identity or equicorrelation estimators proposed by Ledoit and Wolf (2003, 2004), setting the shrinkage parameters to their respective optimal values derived in the aforementioned papers. We also consider the hard and soft thresholding estimators due to Bickel and Levina (2008b) and Rothman, Levina and Zhu (2009). The various fine-tuning parameters required to calculate these estimators are set in accordance with the default values of the R code kindly shared with us by Adam J. Rothman. The performance of our estimator is usually the best in the case of fat-tailed data. In the case of Gaussian data, it is dominated sometimes by thresholding and sometimes by shrinkage, but is never far from the best-performing method. Compare this to the $\mathrm{R}_{2,\mathrm{S}}$ loss of our best-performing competitor, soft thresholding, when $k = 250$ and $\rho = 0.95$: we dominate by a factor of two. In unreported simulations, we experimented with a few skewed distributions and the result was the dominance of our estimator again, suggesting that the thickness of either tail is what may be driving the rankings, regardless of whether it is one or both tails that are thick.

3.2. *Reduction in ill-conditioning.* Moving to analyze the reduction in ill-conditioning, Table 3 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. We find that choosing small m delivers the largest improvements in the conditioning of the estimated variance matrix, and that the gains are massive. The two-step estimator achieves up to 100 times smaller condition number than the sample variance matrix. The improvements are uniform over the different values of ρ.

We found in the previous subsection that the efficiency of the grand-averaging estimator $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ of (3.1) was often the best, compared to the baseline estimator where m is to be chosen optimally. We now see in Table 3 that the price to pay for this increase in precision is a slight increase in ill-conditioning, since the condition number is seen to rise with m.

An attractive feature of our 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 $\widetilde{\boldsymbol{\Sigma}}$ 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. *Data-dependent procedure to choose m.* We next turn to the optimal choice of m in practical applications. One possibility is to use the grand-averaging estimator $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ of (3.1). Another one is to use resampling techniques to make an explicit choice about one value for m, which is what we consider in this subsection. The i.i.d. setup of the previous section (and the moment existence condition) 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

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

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

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

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 $\mathcal{M} := \{m_1, m_2, \ldots, m_M\}$ like before, 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

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

Results are reported in Table 4 for our baseline estimator $\widetilde{\boldsymbol{\Sigma}}_m$ and for its general version $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ of (2.15). We see a very good performance of the suggested bootstrap 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. In the case of the baseline estimator, it is well below 1%. In the case of the more general estimator (where the MSE is lower to start with), the loss goes up to about 4%.

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 elsewhere.

Because our estimator is nonsingular with probability 1 even for $k \geq n$, a case where the traditional estimator of $\boldsymbol{\Sigma}$ is singular, our approach opens up a host of other applications. This is for example the case in longitudinal analysis (like panel data) if one does not wish to impose restrictive assumptions on the covariance structure of the model.

Acknowledgements. We are grateful for the comments received by Valentina Corradi, Oliver Linton, seminar participants at Cambridge University, GREQAM, Lancaster University, LSE, Université de Lausanne, conference participants at the African Econometric Society, Computational and Financial Econometrics, Rimini Workshop in Time Series, Vilnius Conference on Probability Theory and Mathematical Statistics. We thank Adam J.

Rothman for sharing his R code for computing the thresholding estimators. We also thank Simon Burbidge and Matt Harvey for their assistance with using the Imperial College High Performance Computer Cluster where the simulations reported in this paper were run.

TABLE 1. *Estimates of the precision criterion* $\mathrm{R}_{2,S}$.[a]

		$k = 30$								$k = 100$								$k = 250$					
		Gaussian			t(5)					Gaussian			t(5)					Gaussian			t(5)		
		ρ			ρ					ρ			ρ					ρ			ρ		
n	m	0.5	0.75	0.95	0.5	0.75	0.95	n	m	0.5	0.75	0.95	0.5	0.75	0.95	n	m	0.5	0.75	0.95	0.5	0.75	0.95
20	5	8.57	21.2	40.1	13.9	26.8	70.5	50	10	29.1	79.0	138	33.7	91.2	216	100	20	74.5	213	365	81.0	234	558
	10	8.20	18.1	37.2	14.1	25.0	71.9		20	27.2	63.6	124	31.8	79.1	212		40	70.1	175	309	77.4	202	515
	15	9.37	17.8	39.8	14.6	27.0	74.8		30	25.6	59.8	126	31.8	77.1	220		60	68.3	162	311	75.7	192	524
									40	27.7	62.2	132	33.1	80.9	252		80	68.6	161	335	77.0	194	568
	av	8.19	18.4	37.1	12.9	24.4	69.1		av	26.6	62.2	122	31.1	77.0	209		av	68.7	169	306	75.3	196	508
	$\widehat{\boldsymbol{\Sigma}}$	**24.1**	**25.3**	**32.8**	**65.6**	**71.0**	**99.7**		$\widehat{\boldsymbol{\Sigma}}$	**101**	**103**	**117**	**266**	**274**	**337**		$\widehat{\boldsymbol{\Sigma}}$	**313**	**315**	**334**	**800**	**801**	**847**
30	10	7.19	14.3	25.6	9.62	20.6	59.9	100	20	24.6	49.2	69.3	28.3	62.2	135	250	50	60.7	118	144	67.0	147	292
	15	6.94	13.6	24.8	9.54	20.7	60.5		40	21.7	38.9	64.9	26.3	54.5	132		100	53.3	93.0	133	61.7	129	276
	20	7.17	14.1	24.9	10.0	22.0	65.4		60	20.7	37.3	64.6	26.3	53.6	136		150	50.8	88.3	135	59.8	125	278
	25	8.67	16.3	29.8	11.0	23.3	63.8		80	21.1	39.0	66.3	26.5	56.1	142		200	50.3	88.5	141	60.8	128	303
	av	6.93	13.7	24.5	9.25	20.1	58.5		av	21.2	38.5	66.7	25.8	53.6	129		av	52.1	91.8	134	60.6	127	271
	$\widehat{\boldsymbol{\Sigma}}$	**16.3**	**17.1**	**22.5**	**51.1**	**59.3**	**94.1**		$\widehat{\boldsymbol{\Sigma}}$	**51.2**	**52.2**	**59.9**	**139**	**144**	**179**		$\widehat{\boldsymbol{\Sigma}}$	**126**	**127**	**133**	**348**	**357**	**385**
50	10	6.35	11.0	16.5	8.09	15.9	39.0	500	100	44.5	63.6	73.9	52.7	95.6	170	500	100	44.5	63.6	73.9	52.7	95.6	170
	20	5.47	9.16	15.0	7.67	15.1	39.9		100	13.3	18.1	26.8	18.1	31.3	64.6		200	37.5	52.7	70.6	47.9	85.9	162
	30	5.41	9.31	14.9	7.82	15.7	40.1		150	12.7	17.8	26.2	17.7	31.1	65.5		300	35.7	50.8	70.8	46.9	84.0	166
	40	6.03	10.3	16.4	8.71	17.4	40.0		200	12.8	18.8	26.6	18.1	33.1	65.9		400	35.3	51.3	72.4	47.0	85.2	176
	av	5.37	9.08	14.6	7.49	14.9	38.0		av	13.1	17.9	26.2	17.8	31.0	63.1		av	36.9	52.1	70.3	47.3	84.8	160
	$\widehat{\boldsymbol{\Sigma}}$	**9.88**	**10.3**	**13.6**	**33.9**	**37.5**	**54.2**		$\widehat{\boldsymbol{\Sigma}}$	**20.6**	**21.0**	**23.8**	**56.8**	**58.4**	**73.2**		$\widehat{\boldsymbol{\Sigma}}$	**63.2**	**63.6**	**67.4**	**170**	**173**	**188**

[a] Notes: Bold entries refer to the sample variance matrix ($n = m$) and shaded cells report the minimum value over m. The table 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 (2.15). 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 (3.1). All results are based on 1,000 simulations.

Table 2
Comparison of alternative estimators for $n = 100$.

			Gaussian			t(5)	
			ρ			ρ	
k	Estimator	0.5	0.75	0.95	0.5	0.75	0.95
30	Sample covariance	5.01	5.29	6.89	21.9	20.9	24.7
	Our estimator	3.51	5.13	7.35	5.76	9.79	20.4
	Shrinkage identity	3.38	4.76	7.01	6.16	9.93	16.4
	Shrinkage equicorrelation	3.53	4.75	6.96	7.67	10.9	21.6
	Hard thresholding	4.96	6.19	6.89	12.9	16.0	20.9
	Soft thresholding	3.89	5.50	7.06	10.4	14.3	19.6
100	Sample covariance	51.3	52.3	59.3	138	144	181
	Our estimator	21.2	38.3	64.3	25.4	53.7	137
	Shrinkage identity	20.3	37.6	57.0	27.1	61.7	123
	Shrinkage equicorrelation	21.3	37.3	56.5	31.1	63.2	134
	Hard thresholding	30.0	42.3	64.0	40.1	117	164
	Soft thresholding	20.1	35.3	66.8	38.5	84.2	165
250	Sample covariance	315	317	335	963	993	1035
	Our estimator	68.3	168	301	76.6	197	514
	Shrinkage identity	66.1	161	295	79.4	222	588
	Shrinkage equicorrelation	69.6	160	291	93.0	229	575
	Hard thresholding	86.0	144	378	104	333	954
	Soft thresholding	63.0	124	314	103	272	803

Table 3

Average ratio $\widetilde{c}_{n-m}/\widehat{c}_n$ of condition numbers.

		Gaussian			t(5)		
			ρ			ρ	
n	m	0.5	0.75	0.95	0.5	0.75	0.95
$k = 30$							
	10	0.023	0.023	0.028	0.012	0.014	0.020
	20	0.030	0.036	0.047	0.015	0.021	0.031
50	30	0.037	0.048	0.061	0.020	0.029	0.041
	40	0.050	0.063	0.076	0.027	0.039	0.051
	av	0.029	0.035	0.044	0.014	0.019	0.028
$k = 100$							
	50	0.058	0.062	0.078	0.024	0.032	0.052
	100	0.074	0.091	0.118	0.032	0.049	0.080
250	150	0.087	0.112	0.140	0.038	0.060	0.096
	200	0.103	0.133	0.158	0.046	0.075	0.109
	av	0.073	0.089	0.112	0.031	0.047	0.075
$k = 250$							
	100	0.032	0.034	0.042	0.010	0.014	0.025
	200	0.040	0.049	0.066	0.013	0.021	0.041
500	300	0.048	0.061	0.081	0.015	0.027	0.051
	400	0.055	0.073	0.092	0.018	0.033	0.059
	av	0.041	0.048	0.063	0.013	0.021	0.039

Table 4

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

	No averaging			Averaging over 20 samples		
		ρ			ρ	
	0.5	0.75	0.95	0.5	0.75	0.95
Optimal m	20	21	16	25	25	25
Median	24	22	16	24	22	16
Mean	23.5	22.1	16.3	23.6	22.2	16.4
Std. Dev.	1.22	1.61	2.31	1.14	1.80	2.48
Min	20	16	11	19	18	11
10% quantile	22	20	13	22	20	13
25% quantile	23	21	15	23	21	15
75% quantile	24	23	18	25	23	18
90% quantile	25	24	19	25	25	20
Max	27	26	23	28	27	24
Increase in MSE	0.49%	0.05%	0.00	0.37%	0.22%	3.98%

Notes: This table reports results for the bootstrap procedure to choose m in order to minimize the 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.

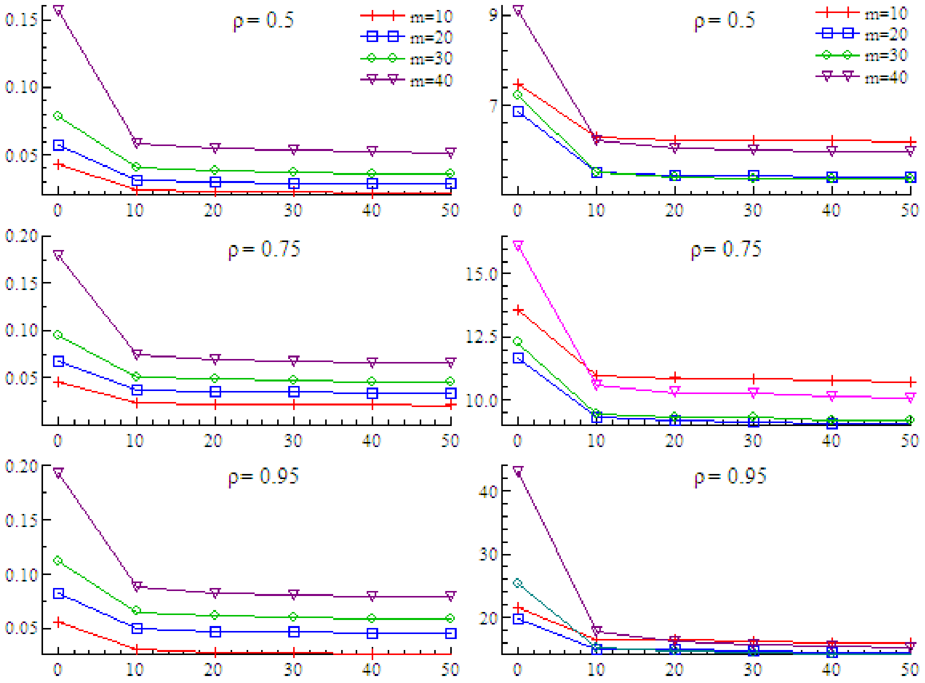

Fig 1. *Ratio of condition numbers* $\widetilde{c}_{n-m}/\widehat{c}_n$ *(left panel) and* $\mathrm{R}_{2,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}}_{m,S})$ *(right panel), averaged over* S *simulations (horizontal axes), for* $k = 30$, $n = 50$, *and* $\boldsymbol{x} \sim \mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$.

References.

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

[2] Anderson, T.W. (1963). Asymptotic Theory for Principal Component Analysis. *Annals of Statistics*, 34, 122–148.

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

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

[5] Brodie, J., De Mol, C., Daubechies, I., Giannone, D. and Loris, I. (2009). Sparse and Stable Markowitz Portfolios. *Proceedings of the National Academy of Science*, 106, 12267–12272.

[6] Davis, A.W. (1977). Asymptotic Theory for Principal Component Analysis: Nonnormal Case. *Australian Journal of Statistics*, 19, 206–212.

[7] DeMiguel, V., Garlappi, L., Nogales, F. and Uppal, R. (2009). A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms. *Management Science*, 55, 798–812.

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

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

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

[11] Fan, J., Zhang, J. and Yu, K. (2008). Asset Allocation and Risk Assessment with Gross Exposure Constraints for Vast Portfolios. Working Paper, Princeton University.

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

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

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

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

[16] 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.

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

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

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

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

[21] Rothman, A.J., Levina, E. and Zhu, J. (2009). Generalized Thresholding of Large Covariance Matrices. *Journal of the American Statistical Association*, 104, 177–186.

[22] 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.

[23] 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.

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

Imperial College London
Imperial College Business School
South Kensington Campus
Exhibition Road
London SW6 2AZ
United Kingdom
E-mail: k.m.abadir@imperial.ac.uk
w.distaso@imperial.ac.uk
fzikes@imperial.ac.uk