Design-free estimation of variance matrices

Karim M. Abadir
Imperial College London

Walter Distaso
Imperial College London

Filip Žikeš
Bank of England

Summary. 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.

Keywords: Variance matrices, ill-conditioning, mean squared error, mean absolute deviations, resampling, U-statistics.

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) propose banding methods to find consistent estimators of variance matrices and their inverse. Other authors resort to thresholding (Bickel and Levina, 2008b, El Karoui, 2008 and Fan, Liao, and Mincheva, 2013) 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) proposed 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 (2012). Fan, Liao, and Mincheva (2011) combine a factor structure with sparsity of the variance matrix.

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 (2003, 2004a,b, 2012). 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 (2003) propose an optimal estimation procedure for the shrinkage parameter, where the chosen metric is the Frobenius norm between the variance and the shrinkage matrix. An alternative approach whereby off-diagonal elements are downweighted towards zero is given in McMurry and Politis (2010) and Politis (2011) in the context of time series. See also an approach to shrinkage via condition-number regularization in Won, Lim, Kim, and Rajaratnam (2013).

In this paper, we introduce a new method to estimate 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 (a subsample), 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 are easy to solve. Moreover, we improve precision further by repeating our procedure over different subsamples, and we show that averaging the resulting estimators leads to a superior performance.

Even though we only use the simple traditional formula for the sample variance matrix in both steps of our basic orthogonalization-estimation procedure, the result is a well-conditioned and precise estimator. Because of the orthogonalization of the data, the resulting estimate is positive definite with probability one, *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. We also show that our simple estimator does very well in comparison to other

existing methods.

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 simplest baseline 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 further guidance on its use in practice. Section 4 concludes. The derivations are collected in the Appendix.

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 properties for the simplest baseline formulation of our estimator, and afterwards we tackle the full version of it as an extension for which the properties are then easily obtained. We describe the optimal choice of two subsampling parameters (one for each step of the baseline orthogonalization-estimation procedure), first in the case of fixed k, then when k expands as n increases.

2.1 The setup and the main idea behind the orthogonalization-estimation procedure

Let $\boldsymbol{\Sigma} := \text{var}(\boldsymbol{x})$ 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)

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

where $\boldsymbol{M}_n := \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{\imath}_n\boldsymbol{\imath}_n'$ is the de-meaning 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 de-meaning 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

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

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 ratio of 1 being the lowest (best numerical) condition. 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 (1) as

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

the last equality following 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 $\widehat{\boldsymbol{p}}_i'\boldsymbol{x}$ did in (2) for all the observations) 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 $\boldsymbol{\Lambda}$: this will be seen

in Proposition 2 (for the condition number) and implied by the estimators' expansions in Propositions 3–4 (for the error norms). As a result, m becomes the only remaining subsampling parameter in question. Propositions 3–4 will show that the precision of the new estimator is optimized by expressing m as a function of n asymptotically. Proposition 5 then extend these results to the case when k varies as n increases, and only then do we consider the alternative definition of consistency as convergence in mean square. These propositions are followed by a concluding discussion of how to calculate the optimal m by resampling in any finite sample, not just asymptotically.

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), 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. We will prove this in the next subsection.

Another advantage of our procedure is that we can estimate the matrix itself as well as any function thereof in one go from the eigenvalue decomposition. The other methods seen in the introduction focus on the variance matrix, and if a function is needed (such as the inverse), one has to make further multivariate calculations to obtain it. This can be 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 simple traditional sample variance estimator in both steps of our procedure.

2.2 Properties of the baseline estimator and its general version

To summarize the procedure in equations, we start by writing

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

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

$$\widehat{\boldsymbol{\Sigma}}_1 := \frac{1}{m} \boldsymbol{X}_1' \boldsymbol{M}_m \boldsymbol{X}_1 = \widehat{\boldsymbol{P}}_1 \widehat{\boldsymbol{\Lambda}}_1 \widehat{\boldsymbol{P}}_1', \tag{4}$$

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

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

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

$$\widetilde{\boldsymbol{\Sigma}} := \widehat{\boldsymbol{P}}\widetilde{\boldsymbol{\Lambda}}\widehat{\boldsymbol{P}}' = \widehat{\boldsymbol{P}}\operatorname{dg}\left(\widehat{\boldsymbol{P}}_1'\widehat{\boldsymbol{\Sigma}}_2\widehat{\boldsymbol{P}}_1\right)\widehat{\boldsymbol{P}}'. \tag{6}$$

Note that we use the simple traditional estimator of variance matrices $\widehat{\operatorname{var}}(\cdot)$ 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, we choose to de-mean $\boldsymbol{X}_2$ 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. Second, by standard statistical analysis, 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, restricting off-diagonal elements to be zero. Third, the estimator $\widetilde{\boldsymbol{\Sigma}}$ is almost surely nonsingular, like the true $\boldsymbol{\Sigma}$, as we now show.

Proposition 1 *For any k, m, n and positive definite $\boldsymbol{\Sigma}$, the estimators $\widetilde{\boldsymbol{\Lambda}}$ and $\widetilde{\boldsymbol{\Sigma}}$ are positive definite with probability 1.*

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 we have used so far.

Proposition 2 *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*

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

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 within the class of estimators $\widetilde{\boldsymbol{\Lambda}}_j$ is minimized with probability 1 by choosing $j/m \to 1$.

The estimator $\widetilde{\boldsymbol{\Lambda}}_j$ differs slightly from the one used in (5) for $j = m$, because of the de-meaning by the whole sample's mean $\overline{\boldsymbol{x}}$ in the proposition, as opposed to $\boldsymbol{X}_2'\boldsymbol{M}_{n-m}\boldsymbol{X}_2$ of (5) de-meaning 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 used to prove the proposition (Anderson, 1963, for the case of $\boldsymbol{x}$ normal and its extension in Davis, 1977), but we conjecture that it is not a necessary condition.

Note the conditions $n-m \to \infty$ and $m \to \infty$, needed for the consistency of the estimator in the sense of its stochastic convergence to the true value. We now turn to the question of 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 $\widetilde{\boldsymbol{\Sigma}}$:

$$\mathrm{R}_l(\widetilde{\boldsymbol{\Sigma}}) := \mathrm{E}(||\operatorname{vec}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})||_l^l), \qquad l = 1, 2, \tag{7}$$

and

$$\mathrm{R}_{l,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}}) := \mathrm{E}(||\operatorname{vech}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})||_l^l), \qquad l = 1, 2, \tag{8}$$

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.

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, depending on the application for which the estimator $\widetilde{\boldsymbol{\Sigma}}$ is to be used; e.g., if we require the same weight to be given to each distinct element in the estimators of $\boldsymbol{\Sigma}$ then the S version is the one to use. At the end of this section, we will discuss the choice of m in finite samples, in which case the criterion's selection makes a difference.

Proposition 3 *As $n-m \to \infty$ and $m \to \infty$, the precision criteria in (7)–(8) 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.*

The case of a scalar matrix $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_k$ is essentially the case of the estimation of one scalar σ^2 that is the same for every variate; which is rare and even more restrictive than assuming that the variates are uncorrelated (diagonal $\boldsymbol{\Sigma}$). In (21) in the proof, 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 choice of m is to take $n - m$ as large as possible to minimize the leading term. This 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. This also shows that the exclusion only applies to scalar matrices but not to the diagonal $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{D}$, since $\boldsymbol{Q}(\sigma^2 \boldsymbol{D})\boldsymbol{Q}' \neq \sigma^2 \boldsymbol{D}$ in general. Special treatment of the scalar-matrix case is not an artefact of our approach: it leads to a singularity in the distribution theory (empty sums arise) and is common to the literature on the asymptotics of sample variance matrices cited after Proposition 2.

Up to now, we constructed our simple estimator $\widetilde{\boldsymbol{\Sigma}}$ by basing $\widehat{\boldsymbol{P}}_1$ on the first m observations 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 (21) as a U-statistic (that is centred around zero by the independence of the subsamples generating $\boldsymbol{Z}_2$ and $\widehat{\boldsymbol{Q}}_{1,ii}$ in (21)) and reduce its magnitude as we will now show. The sampling intuition behind this additional step 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

$$\widetilde{\boldsymbol{\Sigma}}_{m,S} := \frac{1}{S}\sum_{s=1}^{S}\widetilde{\boldsymbol{\Sigma}}_{m,s} \tag{9}$$

denote the average over S different resamples. 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 3, the properties derived earlier for our simpler estimator apply also to our general estimator (9). As for the choice of m, we get the following result to complement Proposition 3.

Proposition 4 *As $n-m \to \infty$ and $m \to \infty$, the precision criteria in (7)–(8) 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.*

The leading term of $\widetilde{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}$ is now $O_p(n^{-1/2})$ independently of m, which we will show numerically 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 baseline estimator $\widetilde{\boldsymbol{\Sigma}}$, as we will now show. Let $\boldsymbol{\Sigma}$ have r distinct eigenvalues, $\lambda_1 > \lambda_2 > \ldots > \lambda_r > 0$ with multiplicities $k_1, k_2, \ldots, k_r$. 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 (21) in the Appendix gives

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

By choosing $m/n \to 0$ with $m/\sqrt{n} \to \infty$ as in Proposition 3, this becomes

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

The resulting sum of two orders is the same as the one obtained in the proof of Proposition 4, but the $O_p(m^{-1})$ term is bigger in the baseline case because the optimal m is such that $m/n \to 0$ in Proposition 3 while $m/n \to \gamma \in (0,1)$ in Proposition 4.

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 3 where we got

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

we can write the expansion for the traditional estimator as

$$\widehat{\boldsymbol{\Lambda}}_i - \lambda_i \boldsymbol{I}_{k_i} \overset{a}{=} n^{-1/2} \widehat{\boldsymbol{Q}}'_{ii} \boldsymbol{P}'_i \boldsymbol{Z} \boldsymbol{P}_i \widehat{\boldsymbol{Q}}_{ii} + n^{-1} \sum_{j \neq i} \lambda_j \widehat{\boldsymbol{R}}'_{ji} \widehat{\boldsymbol{R}}_{ji} - n^{-1} \lambda_i \widehat{\boldsymbol{W}}_{ii}, \tag{10}$$

where $\boldsymbol{Z}$ is normal, like $\boldsymbol{Z}_2$ was for $\widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\boldsymbol{\Sigma}}_{m,q}$, but is now pertaining to the full sample rather than the second subsample only. No resampling-and-averaging will reduce the leading term in (10), because $\widehat{\boldsymbol{Q}}_{ii}$ and $\boldsymbol{Z}$ are (cor)related in general, unlike in the case of our procedure's $\widehat{\boldsymbol{Q}}_{1,ii}$ and $\boldsymbol{Z}_2$ being from the two independent subsamples.[1] Furthermore, this correlation means that the leading term of $\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}$ is not centred around zero when n is finite, unlike in the case of our estimators. This also confirms, from the different angle of error norms, what Proposition 2 implied earlier that we should not reuse data from step 1 in step 2 of our estimators.

Now consider the case where k expands as n increases, and note the following about the setup. First, clearly k is a natural number, which is not dense in $\mathbb{R}_+$ where λ takes values. As a result, when k increases with n, there is no reason to assume that some distinct λ's would have to converge to a common real number. Therefore, in the proposition below, when some of the λ's are distinct, they will remain so as k increases. Second, we tackle the three questions of the definition of consistency, maximum speed at which k can grow, and optimal choice of m when k expands.

We start by addressing the question of how quickly k can be allowed to grow as n increases, a case where one could use a sieve version of our estimator subject to a restriction on the speed of growth of k. For example, the mean vector can be estimated by a sieve as in Antoniadis (1988, Theorem 3.1) and Ackerberg, Chen, and Hahn (2012), then the variance matrix obtained consistently as in Lemma 2 of the latter reference. Using the element-wise definition of convergence employed so far, consistency requires $k = o(n)$ for both estimators $\widetilde{\boldsymbol{\Sigma}}$ and $\widetilde{\boldsymbol{\Sigma}}_{m,q}$. But the more commonly-used convergence mode in the literature on large variance matrices is convergence in mean square ($l = 2$ in our notation), in which case the order under our norms in (7)–(8) would be $k = o(\min\{\sqrt{m}, \sqrt{n-m}\})$ (equivalently, $k = o(\sqrt{m})$ and $k = o(\sqrt{n-m})$) since we are using the traditional sample variance in both

[1] Designing an alternative to $\widetilde{\boldsymbol{\Sigma}}_{m,q}$ that allows sampling with replacement would have led to an estimator 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 also because of the increase in condition numbers that arises from reusing observations (see Proposition 2).

steps of our procedure. To determine the optimal m, we now need the following proposition to complement the previous two.

Proposition 5 *Assume that the positive definite $\boldsymbol{\Sigma}$ is not a scalar matrix. As $n-m \to \infty$ and $m \to \infty$, the precision criteria in (7)–(8), with $l = 2$, are optimized asymptotically:*

1. *for $\widetilde{\boldsymbol{\Sigma}}$ by taking m growing as fast as possible subject to $m/n \to 0$;*
2. *for $\widetilde{\boldsymbol{\Sigma}}_{m,q}$ by taking $m/n \to \gamma$ with $\gamma \in (0,1)$;*

when k is allowed to grow at any rate subsumed by $k = o(\sqrt{n})$.

The result for our simple $\widetilde{\boldsymbol{\Sigma}}$ has changed compared to Proposition 3: we now require m to grow as fast as possible (not just that m grows faster than $\sqrt{n}$) subject to it growing slower than n. On the other hand, the result for our general $\widetilde{\boldsymbol{\Sigma}}_{m,q}$ is unchanged compared to Proposition 4: as the proof shows, we could have allowed $m/n \to 1$ with $n - m \to \infty$, such as $m = n - \log n$, but this would be a case where

$$k = o(\min\{\sqrt{m}, \sqrt{n-m}\}) = o(\sqrt{n-m}) = o(\sqrt{\log n})$$

and hence the rate at which k is allowed to grow is substantially diminished without any efficiency gain in return. (Remember that m is a subsampling parameter that the user chooses, and this selection should not be detrimental to other performances.) This is why we restrict $\gamma \in (0,1)$ and exclude the case $\gamma = 1$, hence allowing k to grow at *any* rate subsumed by $k = o(\sqrt{n})$.

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 estimator, to compare its performance with its most popular competitors, and to automate the choice of m by resampling. In Subsection 3.1, we investigate 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 3.2, we investigate the reduction in the condition

number of our estimator relative to the sample variance matrix benchmark. Subsection 3.3 compares the performance of our estimator to its leading competitors. Finally, Subsection 3.4 investigates the automation of the choice of m by using the bootstrap.

Our simulation design is as follows. 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 by t(5), such that the first four moments exist. The population mean is taken to be zero, without loss of generality because of the location invariance of our estimators. We take $k \in \{30, 100, 250\}$ as the dimension of the random vector $\boldsymbol{x}$, and report three values of n accordingly. 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.

We consider three alternative designs for the true variance matrix $\boldsymbol{\Sigma}$. First, we let it have a Toeplitz structure with typical element given by $\sigma_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, k$, with $\rho \in \{0.0, 0.5, 0.75, 0.95\}$. Note that our estimation method is invariant to rotations and reflections (orthogonal transformations of $\boldsymbol{x}$), hence not affected by changing the implicit ordering that follows from the Toeplitz structure. Second, we take an equicorrelation matrix $\sigma_{ii} = 1$ and $\sigma_{ij} = \rho$ for $i \neq j$, with $\rho \in \{0.0, 0.5, 0.75, 0.95\}$ again. Third, we take a perturbation of the equicorrelation matrix, which we call a "uniform" design: $\sigma_{ii} = 1$ and $\sigma_{ij} = \alpha \mathrm{U}_{(0,1)}$ for $i \neq j$, where $\mathrm{U}_{(0,1)}$ is a standard uniform variate. We take $\alpha \in \{0.0, 0.05, 0.1\}$, as higher values of α lead to violations of positive definiteness in our setup.

Recalling the definition of our general estimator (9), 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 $\widetilde{\boldsymbol{\Sigma}}_{m,S}$. Each curve represents a different value of m. 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. This choice of S was for convenience in our simulations. In practice, S could be chosen alternatively such that

increasing it does not affect the numerical value of $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ to some required digits.

3.1 Estimator's Precision

The results are summarized in Tables 1–3, where the line labelled "av" will be introduced and analyzed in the next paragraph. 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 estimator dominates, except in the case when ρ equals the extreme 0.95 *and* the data are Gaussian. But even in this case, ours dominates as k increases relative to n. In all other cases, the achieved reduction in the mean squared error is very large and is more pronounced for data generated from the fat-tailed Student t distribution and/or smaller ρ. The gains are truly massive in the cases of Toeplitz and uniform designs, e.g., $\mathrm{R}_{2,\mathrm{S}}$ is better for our estimator by a factor of eleven in the case of $\mathrm{t}(5), n = 100, k = 250, \rho = 0.5$ in Table 1 and, for the same $\mathrm{t}(5), n, k$, by a factor of 32 when $\alpha = 0.05$ in Table 3 (the factor even reaching 36 when $n = 50$ and $k = 100$). In the case of a scalar matrix (the same $\boldsymbol{\Sigma}$ for the three tables) and $k = 250$, the result is an improvement factor that ranges between 1361–3155 (increasing with n) in the Gaussian case, and around 180 in the case of $\mathrm{t}(5)$.

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 4 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

$$\widetilde{\boldsymbol{\Sigma}}_{M,S} := \frac{1}{M} \sum_{m \in \mathcal{M}} \widetilde{\boldsymbol{\Sigma}}_{m,S}, \tag{11}$$

where M is the number of elements in the grid $\mathcal{M} := \{m_1, m_2, \ldots, m_M\}$, where $1 < m_1 < \cdots < m_M < n - 1$. Results are reported in the line labelled "av". The performance of $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ is very good in terms of precision. In most cases, $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ is the most precise estimator or close to being so, hence providing an estimator with good overall risk.

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 $\rho = 0.95$.

3.2 Reduction in ill-conditioning

Moving to analyze the reduction in ill-conditioning, Tables 4–6 report 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, but the gains remain even if m increases. These are massive: our estimator achieves up to 100 times smaller condition number than the sample variance matrix. The improvements are uniform over the different values of ρ and α.

We found in the previous subsection that the efficiency of the grand-averaging estimator $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ of (11) was often the best, compared to the baseline estimator where m is to be chosen optimally. We now see that the price to pay for this increase in precision is an occasional slight increase in the condition number (because it rises 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 $\widehat{\boldsymbol{\Sigma}}$ is not positive definite. Unreported simulations for the Toeplitz case 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 $\widehat{\boldsymbol{\Sigma}}$ with $n = 50$.

3.3 Comparison with leading existing estimators

We also compare the performance of our estimator with its most popular competitors, in Tables 7–9 for the MSE-type $\mathrm{R}_{2,\mathrm{S}}$ and Tables 10–12 for condition numbers. We consider the shrinkage towards identity or equicorrelation estimators proposed by Ledoit and Wolf (2004a,b), 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.

When $\boldsymbol{\Sigma}$ is a Toeplitz matrix, Table 7 indicates that 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 in this table, 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 fat-tailed rankings, regardless of whether it is one or both tails that are thick. When large, a Toeplitz matrix is sparse by construction, which gives the best shot to thresholding method. We now turn to the remaining two designs of $\boldsymbol{\Sigma}$ that give shrinking the best shot.

When $\boldsymbol{\Sigma}$ is an equicorrelation matrix, it is natural to expect the method that shrinks towards the true $\boldsymbol{\Sigma}$ to perform better in Table 8. It turns out that this is true when the data are Gaussian, but that shrinking towards an identity matrix is even better as ρ increases towards the singularity region and MSE increases, especially under fat tails. In fact, when ρ is high or when the tails are fat, our estimator often beats shrinking towards an equicorrelation matrix.

When $\boldsymbol{\Sigma}$ is a random perturbation of an equicorrelation matrix, Table 9 reveals that shrinkage towards an equicorrelation is best when the dimensions are large, followed by our method, then shrinking towards an identity matrix. The ranking of those three methods is reversed when the dimensions are smaller. This stable overall performance of our estimator seen in Tables 7–9 confirms the comment about the low overall risk of our estimator at the end of Subsection 3.1.

We now turn to comparing the condition numbers of estimators of $\boldsymbol{\Sigma}$ in Tables 10–12. So far, the analysis of improvements in the condition numbers has been within the class of estimators that are based on the sample variance matrix. Now that we are comparing across different types of estimators, we should keep in mind that the numerical criterion of condition numbers should be read in conjunction with the improvements in precision seen earlier. For an explicit way of combining these two measures, see for example DeMiguel, Martin-Utrera, and Nogales (2013). The user may wish to apply their method by combining

our Tables 7–9 and Tables 10–12 with any chosen combination parameter. Tables 10–12 show that our estimator does best overall, except when ρ or α lead to a nearly singular matrix. We do not report our simulations for thresholding because it does not guarantee the positive definiteness of the estimated matrix (unlike our Proposition 1) and, with nonzero probability in finite samples, it generates negative estimates of the smallest eigenvalues.

In our calculations, we have not used the estimator with optimal m, but rather the grand-averaging estimator $\widetilde{\boldsymbol{\Sigma}}_{M,S}$ of (11). The next subsection will show how to get closer to the optimal m for our estimator. In terms of computation times, the eigenvalue problem is not an onerous one; rather, it is the resampling and averaging that is time consuming. Even so, our numerical procedure is much faster than thresholding. For example, when $\boldsymbol{\Sigma}$ is a Toeplitz matrix with $\rho = 0.75$, $n = 100$, $k = 100$, and Gaussian distribution, the following run times (in seconds) were obtained:

- Sample variance: 0.0028s
- Shrinkage to identity: 0.0029s
- Shrinkage to equicorrelation: 0.0053s
- Our estimator: 0.8247s
- Hard thresholding: 4.0232s
- Soft thresholding: 8.5715s.

Our method and thresholding can be made more or less computationally heavy, depending on the various parameters; e.g., in our case by varying the coarseness of the grid or the number of subsamples over which we average.

3.4 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 (11). 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 (see Davison and Hinkley, 2009). 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

$$\begin{aligned}&\frac{1}{B}\sum_{b=1}^{B}(\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widetilde{\boldsymbol{\Sigma}}_{m,B}))'(\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widetilde{\boldsymbol{\Sigma}}_{m,B})) \\ &+ \left(\frac{1}{B}\sum_{b=1}^{B}\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widehat{\boldsymbol{\Sigma}}_B)\right)' \left(\frac{1}{B}\sum_{b=1}^{B}\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widehat{\boldsymbol{\Sigma}}_B)\right),\end{aligned} \tag{12}$$

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

$$\frac{1}{B}\sum_{b=1}^{B}||\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widehat{\boldsymbol{\Sigma}}_B)||_2^2, \tag{13}$$

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

$$m_B := \underset{m\in\mathcal{M}}{\text{argmin}}\frac{1}{B}\sum_{b=1}^{B}||\text{vech}(\widetilde{\boldsymbol{\Sigma}}_{m,b} - \widehat{\boldsymbol{\Sigma}}_B)||_2^2, \tag{14}$$

whose performance will now be illustrated. Before we do so, recall that the derivations of the previous section excluded the case of a scalar matrix which leads to a singularity in the distributional results. If we were to allow scalar matrices, we need a modification of (12) that takes care of the asymptotics for both types of $\boldsymbol{\Sigma}$.

When $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_k$, the optimal m is 2 and the bias is $O(1/n)$. When $\boldsymbol{\Sigma} \neq \sigma^2\boldsymbol{I}_k$, the asymptotic expansions in the previous section show that the bias is $O\left(1/m\right)$ because

the expectation of the first term in the expansion is zero. Since Proposition 4 gives the optimal m as asymptotically proportional to n, a factor of the form $(m/n)^d$ (with $d > 0$) multiplying the squared bias in (12) should combine both cases and ensure that the bias vanishes asymptotically without altering the rate for the main case of $\boldsymbol{\Sigma} \neq \sigma^2 \boldsymbol{I}_k$. It also has the effect of dampening the finite-sample bias (since $d > 0$ and $m/n < 1$), and the simulations will reveal what d to choose in practice.

Results are reported in Table 13 for our $\widetilde{\boldsymbol{\Sigma}}_{m,S}$ of (9), with $\boldsymbol{\Sigma}$ a Toeplitz matrix. We see a very good performance of the suggested bootstrap procedure. When $d = \frac{1}{4}$, the adjusted bootstrap makes a selection very close to the optimal m and the percentage increase in the mean squared error of our bootstrap-based estimator is minimal. When $\boldsymbol{\Sigma} \neq \sigma^2 \boldsymbol{I}_k$, clearly the best choice would be to use the unmodified bootstrap and set $d = 0$ as in (14), with no adjustment factor needed.

4 Conclusion

In this paper, we provide a novel approach to estimating 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.

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.

The very large reduction in ill-conditioning suggests that our estimator should perform well in cases where matrix-inversion operations are required, as for example in portfolio optimization problems. However, the simple sample variance formula that we use in both steps of our procedure optimizes the LS criterion, which is not optimal for estimating for example the inverse of a variance (the precision matrix). This is a general unsolved problem

in statistical theory: there is no explicit solution for optimizing even simple loss criteria except in some restricted cases such as normality; e.g., see Krishnamoorthy and Gupta (1989). Here, one can do the following. Once the data have been asymptotically-orthogonalized in the first step, the estimation in the second step can then be easily done one dimension at a time and a chosen univariate loss criterion optimized numerically. Preliminary results indicate that this approach and its application to empirical problems works very well, but a full investigation is beyond the scope of this paper. An alternative approach is obtainable from the nonlinear shrinkage of Ledoit and Wolf (2012) as follows. Our variates are asymptotically orthogonalized as a result of the first step, and the transformed variates eventually have a diagonal variance matrix. Linear shrinkage of its estimate (by a scalar) would not add much to the second step, but nonlinear shrinkage is a promising venue to explore because it shrinks various eigenvalues to a different extent.

Acknowledgements

We are grateful for the comments received by Yacine Aït-Sahalia, Valentina Corradi, Sir David R. Cox, Axel Gandy, Oliver Linton, Michael Stephens, the referees, seminar participants at Cambridge University, Czech Academy of Sciences (Prague), GREQAM, Lancaster University, LSE, Tinbergen Institute, Université de Lausanne, University of York, conference participants at the African Econometric Society (Cairo), Computational and Financial Econometrics (London), Exeter Workshop on Econometrics and its Applications in honour of James Davidson, International Society for Non-Parametric Statistics (Chalkidiki), Lancaster-Manchester-Liverpool Workshop, Rimini Workshop in Time Series, Society for Nonlinear Dynamics and Econometrics (Istanbul), 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. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Bank of England or any of its committees. This research is supported by the ESRC grants R000239538, RES000230176, RES062230311 and RES062230790. The Ox code for this estimator is

available at http://bit.ly/1cqMpUR.

Appendix: Proofs

Proof of Proposition 1. Because of the use of dg in (5), the eigenvalues of $\widetilde{\boldsymbol{\Sigma}}$ are the diagonal elements $\widetilde{\lambda}_i$ $(i = 1, \ldots, k)$ of $\widetilde{\boldsymbol{\Lambda}}$. Since $\text{var}(\boldsymbol{x})$ is positive definite and $\widetilde{\lambda}_i$ is the sample variance of a nonzero (the columns of $\widehat{\boldsymbol{P}}_1$ are nonzero) linear combination of $\boldsymbol{x}$, we get $\widetilde{\lambda}_i > 0$ almost surely. □

Proof of Proposition 2. For $m > j + 2$,

$$\begin{aligned} \widetilde{\boldsymbol{\Lambda}}_j &= \frac{1}{n-j}\,\text{dg}\left(\widehat{\boldsymbol{P}}_1' \sum_{i=j+1}^{m} \boldsymbol{y}_i \boldsymbol{y}_i' \widehat{\boldsymbol{P}}_1\right) + \frac{1}{n-j}\,\text{dg}\left(\widehat{\boldsymbol{P}}_1' \sum_{i=m+1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i' \widehat{\boldsymbol{P}}_1\right) \qquad (15) \\ &= \frac{m-j}{n-j}\,\text{dg}\,(\boldsymbol{S}_j) + \frac{n-m}{n-j}\widetilde{\boldsymbol{\Lambda}}_m, \end{aligned}$$

which is a weighted average of $\text{dg}\,(\boldsymbol{S}_j)$ 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 (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 the two steps of 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 (15), the two limiting matrices for the components in (15) 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 $\delta_n := \sqrt{2 \log(\log(n))/n}$ and ω_i^2/n as the asymptotic variance (which exists by assumption) of the estimator of λ_i. Writing c for $c_\infty = \lambda_1/\lambda_k$,

$$c_n = \frac{\lambda_1 + \omega_1 \delta_n}{\lambda_k - \omega_k \delta_n} = \left(c + \frac{\omega_1 \delta_n}{\lambda_k}\right)\left(1 + \frac{\omega_k \delta_n}{\lambda_k} + O(\delta_n^2)\right) = c + \frac{\omega_1 + c\omega_k}{\lambda_k}\delta_n + O(\delta_n^2). \quad (16)$$

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 (15) and dropping the remainder terms, we get the asymptotic condition number of $\widetilde{\boldsymbol{\Lambda}}_j$ 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 $m \to \infty$. □

Proof of Proposition 3. Under the i.i.d. assumption and the existence of fourth moments, $\widehat{\boldsymbol{\Sigma}}_2$ satisfies the CLT

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

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

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

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, $\lambda_1 > \lambda_2 > \ldots > \lambda_r > 0$ with multiplicities $k_1, k_2, \ldots, k_r$. (One may also take the diagonal elements of $\boldsymbol{P}_\cdot$ to be positive for uniqueness considerations, and similarly for other diagonalizations.) Now partition the matrices $\boldsymbol{U}_1$ and $\widehat{\boldsymbol{Q}}_1$ into submatrices with $k_1, \ldots, k_r$ rows and columns as $\boldsymbol{U}_1 = (\boldsymbol{U}_{1,ij})$ and $\widehat{\boldsymbol{Q}}_1 = (\widehat{\boldsymbol{Q}}_{1,ij})$, where $i, j = 1, \ldots, r$. Also, partition $\widehat{\boldsymbol{P}}_1 = (\widehat{\boldsymbol{P}}_{1,i})$ and $\boldsymbol{P} = (\boldsymbol{P}_i)$ as blocks of $k_1, \ldots, k_r$ 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 $\widetilde{\boldsymbol{\Lambda}} = \operatorname{diag}(\ldots, \widetilde{\boldsymbol{\Lambda}}_i, \ldots)$, where $\widetilde{\boldsymbol{\Lambda}}_i$ is now the $k_i \times k_i$ matrix corresponding to the eigenvalue λ_i with multiplicity k_i (not the same $\widetilde{\boldsymbol{\Lambda}}_i$ as in Proposition 2). Then, by direct substitution, we get the leading terms of the asymptotic expansion

$$\begin{aligned}
\widetilde{\boldsymbol{\Lambda}}_i &= \operatorname{dg}(\widehat{\boldsymbol{P}}_{1,i}'\widehat{\boldsymbol{\Sigma}}_2\widehat{\boldsymbol{P}}_{1,i}) \\
&\overset{a}{=} \operatorname{dg}((\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} + m^{-1/2}\sum_{j\neq i}\boldsymbol{P}_j\widehat{\boldsymbol{R}}_{1,ji})'(\boldsymbol{\Sigma} + (n-m)^{-1/2}\boldsymbol{Z}_2)(\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} + m^{-1/2}\sum_{j\neq i}\boldsymbol{P}_j\widehat{\boldsymbol{R}}_{1,ji})) \\
&= \operatorname{dg}(\widehat{\boldsymbol{Q}}_{1,ii}'\boldsymbol{P}_i'\boldsymbol{\Sigma}\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} + (n-m)^{-1/2}\widehat{\boldsymbol{Q}}_{1,ii}'\boldsymbol{P}_i'\boldsymbol{Z}_2\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} \qquad (17) \\
&\quad + m^{-1/2}\sum_{j\neq i}\widehat{\boldsymbol{R}}_{1,ji}'\boldsymbol{P}_j'\boldsymbol{\Sigma}\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} + m^{-1/2}(n-m)^{-1/2}\sum_{j\neq i}\widehat{\boldsymbol{R}}_{1,ji}'\boldsymbol{P}_j'\boldsymbol{Z}_2\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii} \qquad (18) \\
&\quad + m^{-1/2}\sum_{j\neq i}\widehat{\boldsymbol{Q}}_{1,ii}'\boldsymbol{P}_i'\boldsymbol{\Sigma}\boldsymbol{P}_j\widehat{\boldsymbol{R}}_{1,ji} + m^{-1/2}(n-m)^{-1/2}\sum_{j\neq i}\widehat{\boldsymbol{Q}}_{1,ii}'\boldsymbol{P}_i'\boldsymbol{Z}_2\boldsymbol{P}_j\widehat{\boldsymbol{R}}_{1,ji} \qquad (19) \\
&\quad + m^{-1}\sum_{j\neq i}\sum_{l\neq i}\widehat{\boldsymbol{R}}_{1,ji}'\boldsymbol{P}_j'\boldsymbol{\Sigma}\boldsymbol{P}_l\widehat{\boldsymbol{R}}_{1,li} + m^{-1}(n-m)^{-1/2}\sum_{j\neq i}\sum_{l\neq i}\widehat{\boldsymbol{R}}_{1,ji}'\boldsymbol{P}_j'\boldsymbol{Z}_2\boldsymbol{P}_l\widehat{\boldsymbol{R}}_{1,li}). \qquad (20)
\end{aligned}$$

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 (18) and (19) is zero, and the first double sum in (20) can be written as a single sum. The second terms in (18), (19) and (20) will always be of smaller order than the remaining terms, so we can drop them. We are therefore left with

$$\widetilde{\boldsymbol{\Lambda}}_i - \lambda_i\boldsymbol{I}_{k_i} \overset{a}{=} (n-m)^{-1/2}\operatorname{dg}(\widehat{\boldsymbol{Q}}_{1,ii}'\boldsymbol{P}_i'\boldsymbol{Z}_2\boldsymbol{P}_i\widehat{\boldsymbol{Q}}_{1,ii}) + m^{-1}\sum_{j\neq i}\lambda_j\operatorname{dg}(\widehat{\boldsymbol{R}}_{1,ji}'\widehat{\boldsymbol{R}}_{1,ji}) - m^{-1}\lambda_i\operatorname{dg}(\widehat{\boldsymbol{W}}_{1,ii}), \qquad (21)$$

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 $n - m \to \infty$ (e.g., $m = n - \log n$), 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 $\widetilde{\boldsymbol{\Lambda}}$. □

Proof of Proposition 4. 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 (21) to

$$\begin{aligned} \widetilde{\boldsymbol{\Lambda}}_i - \lambda_i \boldsymbol{I}_{k_i} \overset{a}{=} \; & (n-m)^{-1/2} \operatorname{dg}(\boldsymbol{Q}'_{ii}\boldsymbol{P}'_i\boldsymbol{Z}_2\boldsymbol{P}_i\boldsymbol{Q}_{ii}) + ((n-m)m)^{-1/2}\,\boldsymbol{\Delta} \\ & + m^{-1}\sum_{j\neq i}\lambda_j \operatorname{dg}(\widehat{\boldsymbol{R}}'_{1,ji}\widehat{\boldsymbol{R}}_{1,ji}) - m^{-1}\lambda_i \operatorname{dg}(\widehat{\boldsymbol{W}}_{1,ii}), \end{aligned} \tag{22}$$

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 $i_1, \ldots, i_{n-m}$ distinct integers taken from $\{1, \ldots, n\}$, and

$$(n-m)^{-1}\operatorname{var}(s_1 + \cdots + s_l + c) \propto l/(n-m)$$

where $c := (n-m)^{-1/2}\operatorname{E}(s_{l+1} + \cdots + s_{n-m})$. Then, the variance of our U-statistic is

proportional to

$$\binom{n}{m}^{-1} \sum_{l=1}^{n-m} \binom{n-m}{l} \binom{m}{n-m-l} \frac{l}{n-m} = \frac{n-m}{n} \tag{23}$$

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 $\operatorname{var}(y) = \operatorname{var}_q(\mathrm{E}_{z|q}\, y) + \mathrm{E}_q \operatorname{var}_{z|q}(y) = \mathrm{E}_q \operatorname{var}_z(y)$) 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 (22) 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 (23). 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 (21) leads to the same optimal m. □

Proof of Proposition 5. For the baseline estimator $\widetilde{\boldsymbol{\Sigma}}$, consider again the orders of magnitude in (17)–(20), recalling that $\boldsymbol{P}, \widehat{\boldsymbol{Q}}_1$ are orthogonal (rotation/reflection matrices) and that $\widehat{\boldsymbol{R}}'_{1,ji}\widehat{\boldsymbol{R}}_{1,ji} = O_p(1)$. The first term of (20) is now $O_p(k/m)$ instead of $O_p(1/m)$ and, when $k = o(\sqrt{m})$, it dominates the second terms in (18)–(19). As for the final term, it is dominated by the $O_p((n-m)^{-1/2})$ term in (17). The result is

$$\widetilde{\boldsymbol{\Lambda}}_i - \lambda_i \boldsymbol{I}_{k_i} = o_p(m^{-1/2}) + O_p((n-m)^{-1/2})$$

whose norm is minimized by taking m growing as fast as possible such that $m/n \to 0$. Unlike before, $O_p((n-m)^{-1/2})$ is not necessarily the leading term anymore when k grows with n. Note that it is suboptimal to take $m/n \to \gamma \neq 0$, as it makes the leading term

larger. The same optimal m arises when imposing in addition that $k = o(\sqrt{n-m})$, and we get $k = o(\sqrt{n})$.

Similarly, for the general estimator $\widetilde{\boldsymbol{\Sigma}}_{m,q}$, (22) implies

$$\widetilde{\boldsymbol{\Lambda}}_i - \lambda_i \boldsymbol{I}_{k_i} = O_p(k/m) + O_p(n^{-1/2})$$

where we have dropped the $O_p(((n-m)m)^{-1/2})$ term which is now dominated. Consider the case of $m/n \to 1$ with $n - m \to \infty$, such as $m = n - \log n$. This is a case where

$$k = o(\min\{\sqrt{m}, \sqrt{n-m}\}) = o(\sqrt{n-m})$$

and hence the rate at which k is allowed to grow is substantially diminished without any efficiency gain in return since the leading $O_p(n^{-1/2})$ term is unaffected. If we exclude this case, $k = o(\sqrt{m})$ and the optimal m is $m/n \to \gamma$ with $\gamma \in (0,1)$, hence $k = o(\sqrt{n})$. □

References

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

Ackerberg, D., Chen, X. and Hahn, J. (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics*, 94, 481–498.

Anderson, T.W. (1963). Asymptotic theory for principal component analysis. *Annals of Statistics*, 34, 122–148.

Antoniadis, A. (1988). Parametric estimation for the mean of a Gaussian process by the method of sieves. *Journal of Multivariate Analysis*, 26, 1–15.

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.

Davis, A.W. (1977). Asymptotic theory for principal component analysis: non-normal case. *Australian Journal of Statistics*, 19, 206–212.

Davison, A. C. and Hinkley, D. V. (2009). *Bootstrap Methods and Their Application*. Cambridge University Press, New-York.

DeMiguel, V., Martin-Utrera, A. and Nogales, F.J. (2013). Size matters: optimal calibration of shrinkage estimators for for portfolio selection. *Journal of Banking & Finance*, 37, 3018–3034.

El Karoui, N. (2008). Operator norm consistent estimation of a large dimensional sparse covariance matrices. *Annals of Statistics*, 36, 2717–2756.

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., Liao Y. and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, 39, 3320–3356.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. Forthcoming, *Journal of Royal Statistical Society B*.

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.

Krishnamoorthy, K. and Gupta, A.K. (1989). Improved minimax estimation of a normal precision matrix. *Canadian Journal of Statistics*, 17, 91–102.

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. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *Annals of Statistics*, 40, 694–726.

Ledoit, O. and Wolf, M. (2003). 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. (2004a). A well-conditioned estimator for large dimensional covariance matrices. *Journal of Multivariate Analysis*, 88, 365–411.

Ledoit, O. and Wolf, M. (2004b). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, 31, 1–22.

Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, 40, 1024–1060.

McMurry, T. and Politis, D.N. (2010). Banded and tapered estimates of autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, 31, 471–482 [Corrigendum (2012), 33, forthcoming].

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

Politis, D.N. (2011). Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices. *Econometric Theory*, 27, 703–744.

Rothman, A.J., Bickel, P.J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2, 494–515.

Rothman, A.J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104, 177–186.

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.

Won, J.-H., Lim, J., Kim, S.-J. and Rajaratnam, B. (2013). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society B*, 75, 427–450.

Wu, W.B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 94, 1–17.

Table 1: Estimates of the precision criterion $R_{2,S}$, Toeplitz design. [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	0.5	0.75	0.95	0	0.5	0.75	0.95	n	m	0	0.5	0.75	0.95	0	0.5	0.75	0.95	n	m	0	0.5	0.75	0.95	0	0.5	0.75	0.95		
20	5	0.36	8.57	21.2	40.1	2.88	13.9	26.8	70.5	50	10	0.25	29.1	79.0	138	4.37	33.7	91.2	216	100	20	0.23	74.5	213	365	5.04	81.0	234	558		
	10	0.62	8.20	18.1	37.2	3.07	14.1	25.0	71.9		20	0.35	27.2	63.6	124	4.09	31.8	79.1	212		40	0.32	70.1	175	309	5.37	77.4	202	515		
	15	1.77	9.37	17.8	39.8	4.32	14.6	27.0	74.8		30	0.61	25.6	59.8	126	5.42	31.8	77.1	220		60	0.53	68.3	162	311	5.96	75.7	192	524		
											40	1.74	27.7	62.2	132	7.64	33.1	80.9	252		80	1.37	68.6	161	335	7.78	77.0	194	568		
	av	0.60	8.19	18.4	37.1	2.73	12.9	24.4	69.1		av	0.37	26.6	62.2	122	4.39	31.1	77.0	209		av	0.24	68.7	169	306	4.79	75.3	196	508		
	$\widehat{\boldsymbol{\Sigma}}$	**23.6**	**24.1**	**25.3**	**32.8**	**62.5**	**65.6**	**71.0**	**99.7**		$\widehat{\boldsymbol{\Sigma}}$	**101**	**101**	**103**	**117**	**299**	**266**	**274**	**337**		$\widehat{\boldsymbol{\Sigma}}$	**313**	**313**	**315**	**334**	**835**	**800**	**801**	**847**		
30	10	0.25	7.19	14.3	25.6	2.08	9.62	20.6	59.9	100	20	0.11	24.6	49.2	69.3	1.70	28.3	62.2	135	250	50	0.08	60.7	118	144	1.92	67.0	147	292		
	15	0.36	6.94	13.6	24.8	2.16	9.54	20.7	60.5		40	0.15	21.7	38.9	64.9	1.76	26.3	54.5	132		100	0.11	53.3	93.0	133	1.93	61.7	129	276		
	20	0.65	7.17	14.1	24.9	2.56	10.0	22.0	65.4		60	0.25	20.7	37.3	64.6	1.85	26.3	53.6	136		150	0.18	50.8	88.3	135	2.17	59.8	125	278		
	25	1.43	8.67	16.3	29.8	3.52	11.0	23.3	63.8		80	0.68	21.1	39.0	66.3	2.64	26.5	56.1	142		200	0.43	50.3	88.5	141	2.99	60.8	128	303		
	av	0.31	6.93	13.7	24.5	1.89	9.25	20.1	58.5		av	0.13	21.2	38.5	66.7	1.57	25.8	53.6	129		av	0.06	52.1	91.8	134	1.79	60.6	127	271		
	$\widehat{\boldsymbol{\Sigma}}$	**16.0**	**16.3**	**17.1**	**22.5**	**43.1**	**51.1**	**59.3**	**94.1**		$\widehat{\boldsymbol{\Sigma}}$	**50.9**	**51.2**	**52.2**	**59.9**	**126**	**139**	**144**	**179**		$\widehat{\boldsymbol{\Sigma}}$	**126**	**126**	**127**	**133**	**329**	**348**	**357**	**385**		
50	10	0.11	6.35	11.0	16.5	1.35	8.09	15.9	39.0	250	50	0.04	15.8	21.4	29.3	0.81	19.8	34.4	65.5	500	100	0.04	44.5	63.6	73.9	1.08	52.7	95.6	170		
	20	0.15	5.47	9.16	15.0	1.51	7.67	15.1	39.9		100	0.05	13.3	18.1	26.8	0.84	18.1	31.3	64.6		200	0.05	37.5	52.7	70.6	1.18	47.9	85.9	162		
	30	0.27	5.41	9.31	14.9	1.58	7.82	15.7	40.1		150	0.09	12.7	17.8	26.2	0.91	17.7	31.1	65.5		300	0.08	35.7	50.8	70.8	1.38	46.9	84.0	166		
	40	0.69	6.03	10.3	16.4	2.17	8.71	17.4	40.0		200	0.20	12.8	18.8	26.6	1.18	18.1	33.1	65.9		400	0.19	35.3	51.3	72.4	1.72	47.0	85.2	176		
	av	0.16	5.37	9.08	14.6	1.29	7.49	14.9	38.0		av	0.04	13.1	17.9	26.2	0.76	17.8	31.0	63.1		av	0.02	36.9	52.1	70.3	1.04	47.3	84.8	160		
	$\widehat{\boldsymbol{\Sigma}}$	**9.72**	**9.88**	**10.3**	**13.6**	**28.2**	**33.9**	**37.5**	**54.2**		$\widehat{\boldsymbol{\Sigma}}$	**20.5**	**20.6**	**21.0**	**23.8**	**55.3**	**56.8**	**58.4**	**73.2**		$\widehat{\boldsymbol{\Sigma}}$	**63.1**	**63.2**	**63.6**	**67.4**	**187**	**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 (9). 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 (11). All results are based on 1,000 simulations.

Table 2: Estimates of the precision criterion $R_{2,S}$, equicorrelation design. [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	0.5	0.75	0.95	0	0.5	0.75	0.95	n	m	0	0.5	0.75	0.95	0	0.5	0.75	0.95	n	m	0	0.5	0.75	0.95	0	0.5	0.75	0.95
20	5	0.36	30.5	40.2	45.3	2.88	53.8	95.4	132	50	10	0.25	122	158	203	4.37	282	503	593	100	20	0.23	345	509	675	5.04	981	1498	1892
	10	0.62	25.5	37.0	45.6	3.07	52.5	108	138		20	0.35	109	154	206	4.09	271	492	573		40	0.32	328	501	662	5.37	1014	1616	1964
	15	1.77	27.4	41.7	53.5	4.32	52.3	96.0	124		30	0.61	110	161	216	5.42	271	462	563		60	0.53	334	529	712	5.96	999	1562	1864
											40	1.74	120	178	247	7.64	274	519	800		80	1.37	354	564	762	7.78	1322	2094	2290
	av	0.60	26.4	37.5	44.5	2.73	50.5	91.2	124		av	0.37	110	154	202	4.39	262	480	579		av	0.24	327	498	659	4.79	953	1482	1800
	$\widehat{\boldsymbol{\Sigma}}$	**23.6**	**29.0**	**36.6**	**44.0**	**62.5**	**79.9**	**118**	**138**		$\widehat{\boldsymbol{\Sigma}}$	**101**	**126**	**156**	**200**	**299**	**389**	**595**	**603**		$\widehat{\boldsymbol{\Sigma}}$	**313**	**394**	**510**	**663**	**835**	**1255**	**1609**	**1861**
30	10	0.25	17.6	24.4	30.0	2.08	39.2	76.2	88.7	100	20	0.11	55.5	75.7	102	1.70	144	211	293	250	50	0.08	132	195	240	1.92	390	664	922
	15	0.36	16.9	24.4	30.6	2.16	37.8	81.2	87.3		40	0.15	53.3	74.9	100	1.76	142	214	286		100	0.11	129	195	244	1.93	378	652	898
	20	0.65	17.3	25.4	32.3	2.56	39.8	84.6	86.3		60	0.25	55.1	77.7	104	1.85	145	221	324		150	0.18	135	208	261	2.17	383	659	896
	25	1.43	19.1	29.2	38.8	3.52	40.1	78.2	92.3		80	0.68	57.4	87.2	118	2.64	148	232	299		200	0.43	142	222	286	2.99	406	701	961
	av	0.31	17.1	24.1	29.5	1.89	37.3	75.6	82.5		av	0.13	53.3	74.8	99.6	1.57	138	212	285		av	0.06	130	195	242	1.79	374	639	873
	$\widehat{\boldsymbol{\Sigma}}$	**16.0**	**19.6**	**24.1**	**29.5**	**43.1**	**55.9**	**91.7**	**89.6**		$\widehat{\boldsymbol{\Sigma}}$	**50.9**	**64.1**	**76.6**	**98.7**	**126**	**181**	**229**	**296**		$\widehat{\boldsymbol{\Sigma}}$	**126**	**158**	**198**	**237**	**329**	**476**	**684**	**890**
50	10	0.11	10.9	14.8	18.0	1.35	27.2	42.4	56.4	250	50	0.04	21.4	30.9	38.1	0.81	61.2	97.3	144	500	100	0.04	64.2	95.2	118	1.08	196	301	391
	20	0.15	10.1	14.6	18.1	1.51	26.5	44.2	58.1		100	0.05	21.1	31.3	38.8	0.84	59.7	98.8	143		200	0.05	64.5	96.6	120	1.18	198	315	413
	30	0.27	10.1	14.9	19.1	1.58	26.5	46.7	59.0		150	0.09	21.2	31.7	39.2	0.91	61.9	105	154		300	0.08	65.2	99.3	126	1.38	218	362	486
	40	0.69	11.2	17.1	22.3	2.17	31.2	46.3	74.0		200	0.20	23.0	35.0	46.8	1.18	66.9	114	209		400	0.19	69.5	108	139	1.72	215	363	507
	av	0.16	10.0	14.4	17.7	1.29	25.6	42.2	55.7		av	0.04	21.0	30.7	38.0	0.76	60.0	98.6	149		av	0.02	63.7	94.9	118	1.04	195	309	404
	$\widehat{\boldsymbol{\Sigma}}$	**9.72**	**11.9**	**14.7**	**17.7**	**28.2**	**36.3**	**49.4**	**58.6**		$\widehat{\boldsymbol{\Sigma}}$	**20.5**	**25.4**	**31.2**	**37.8**	**55.3**	**76.0**	**103**	**141**		$\widehat{\boldsymbol{\Sigma}}$	**63.1**	**78.8**	**98.1**	**117**	**187**	**246**	**319**	**391**

[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 (9). 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 (11). All results are based on 1,000 simulations.

Table 3: Estimates of the precision criterion $R_{2,S}$, uniform design. [a]

		$k=30$								$k=100$								$k=250$							
		Gaussian			t(5)					Gaussian			t(5)					Gaussian			t(5)				
		α			α					α			α					α			α				
n	m	0	0.05	0.1	0	0.05	0.1	n	m	0	0.05	0.1	0	0.05	0.1	n	m	0	0.05	0.1	0	0.05	0.1		
20	5	0.36	0.71	1.71	2.88	3.27	4.30	50	10	0.25	4.22	14.1	4.37	8.72	20.1	100	20	0.23	21.5	63.4	5.04	30.8	98.3		
	10	0.62	0.97	1.96	3.07	3.51	4.55		20	0.35	4.21	12.8	4.09	8.56	19.7		40	0.32	19.1	54.3	5.37	30.8	94.4		
	15	1.77	2.13	3.12	4.32	4.63	5.69		30	0.61	4.41	12.5	5.42	9.33	20.3		60	0.53	18.4	52.3	5.96	30.9	91.5		
									40	1.74	5.51	13.5	7.64	11.0	21.9		80	1.37	18.9	52.6	7.78	33.3	92.0		
	av	0.60	0.94	1.93	2.73	3.12	4.16		av	0.37	4.19	12.6	4.39	8.48	19.5		av	0.24	18.6	53.5	4.79	30.0	91.0		
	$\widehat{\boldsymbol{\Sigma}}$	**23.6**	**23.6**	**23.7**	**62.5**	**62.7**	**62.7**		$\widehat{\boldsymbol{\Sigma}}$	**101**	**101**	**101**	**299**	**306**	**305**		$\widehat{\boldsymbol{\Sigma}}$	**313**	**314**	**315**	**835**	**948**	**950**		
30	10	0.25	0.59	1.54	2.08	2.47	3.54	100	20	0.11	3.77	11.0	1.70	5.89	16.6	250	50	0.08	14.9	40.1	1.92	26.3	81.3		
	15	0.36	0.71	1.65	2.16	2.58	3.66		40	0.15	3.51	9.36	1.76	5.95	16.1		100	0.11	12.8	35.2	1.93	25.5	70.6		
	20	0.65	1.00	1.94	2.56	3.05	4.12		60	0.25	3.45	8.99	1.85	6.16	15.8		150	0.18	12.4	33.7	2.17	24.8	64.0		
	25	1.43	1.79	2.73	3.52	3.67	4.75		80	0.68	3.81	9.32	2.64	6.92	16.2		200	0.43	12.5	33.5	2.99	25.5	61.6		
	av	0.31	0.66	1.60	1.89	2.28	3.34		av	0.13	3.39	9.18	1.57	5.74	15.5		av	0.06	12.6	34.5	1.79	24.6	65.7		
	$\widehat{\boldsymbol{\Sigma}}$	**16.0**	**16.0**	**16.0**	**43.1**	**43.9**	**44.0**		$\widehat{\boldsymbol{\Sigma}}$	**50.9**	**51.0**	**51.1**	**126**	**139**	**139**		$\widehat{\boldsymbol{\Sigma}}$	**126**	**126**	**126**	**329**	**364**	**365**		
50	10	0.11	0.45	1.35	1.35	1.96	3.05	250	50	0.04	2.79	6.77	0.81	4.67	13.6	500	100	0.04	10.6	28.5	1.08	23.5	63.4		
	20	0.15	0.49	1.34	1.51	1.91	3.01		100	0.05	2.32	5.97	0.84	4.55	12.2		200	0.05	9.48	25.7	1.18	21.6	50.8		
	30	0.27	0.61	1.43	1.58	2.08	3.18		150	0.09	2.24	5.86	0.91	4.56	11.5		300	0.08	9.30	25.8	1.38	19.9	46.0		
	40	0.69	1.03	1.85	2.17	2.84	3.95		200	0.20	2.33	5.94	1.18	4.80	11.2		400	0.19	9.35	25.3	1.72	18.6	42.9		
	av	0.16	0.50	1.33	1.29	1.84	2.92		av	0.04	2.27	5.91	0.76	4.43	11.5		av	0.02	9.40	25.6	1.04	20.1	47.2		
	$\widehat{\boldsymbol{\Sigma}}$	**9.72**	**9.71**	**9.73**	**28.2**	**30.5**	**30.3**		$\widehat{\boldsymbol{\Sigma}}$	**20.5**	20.5	**20.6**	**55.3**	**58.8**	**59.1**		$\widehat{\boldsymbol{\Sigma}}$	**63.1**	**63.2**	**63.3**	**187**	**173**	**175**		

[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 (9). 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 (11). All results are based on 1,000 simulations.

Table 4: Average ratio $\widetilde{c}_{n-m}/\widehat{c}_n$ of condition numbers, Toeplitz design.

		Gaussian				t(5)			
		ρ				ρ			
n	m	0	0.5	0.75	0.95	0	0.5	0.75	0.95
$k = 30$									
	10	0.028	0.023	0.023	0.028	0.014	0.012	0.014	0.020
	20	0.030	0.030	0.036	0.047	0.015	0.015	0.021	0.031
50	30	0.034	0.037	0.048	0.061	0.017	0.020	0.029	0.041
	40	0.045	0.050	0.063	0.076	0.023	0.027	0.039	0.051
	av	0.028	0.029	0.035	0.044	0.014	0.014	0.019	0.028
$k = 100$									
	50	0.063	0.058	0.062	0.078	0.023	0.024	0.032	0.052
	100	0.064	0.074	0.091	0.118	0.024	0.032	0.049	0.080
250	150	0.067	0.087	0.112	0.140	0.026	0.038	0.060	0.096
	200	0.075	0.103	0.133	0.158	0.030	0.046	0.075	0.109
	av	0.062	0.073	0.089	0.112	0.023	0.031	0.047	0.075
$k = 250$									
	100	0.034	0.032	0.034	0.042	0.009	0.010	0.014	0.025
	200	0.035	0.040	0.049	0.066	0.009	0.013	0.021	0.041
500	300	0.036	0.048	0.061	0.081	0.010	0.015	0.027	0.051
	400	0.039	0.055	0.073	0.092	0.011	0.018	0.033	0.059
	av	0.034	0.041	0.048	0.063	0.009	0.013	0.021	0.039

Table 5: Average ratio $\widetilde{c}_{n-m}/\widehat{c}_n$ of condition numbers, equicorrelation design.

		Gaussian				t(5)			
		ρ				ρ			
n	m	0	0.5	0.75	0.95	0	0.5	0.75	0.95
$k = 30$									
	10	0.028	0.059	0.065	0.068	0.014	0.038	0.044	0.047
	20	0.030	0.068	0.072	0.073	0.015	0.044	0.049	0.051
50	30	0.034	0.075	0.079	0.080	0.017	0.050	0.054	0.056
	40	0.045	0.090	0.094	0.095	0.023	0.061	0.066	0.068
	av	0.028	0.066	0.070	0.071	0.014	0.041	0.045	0.047
$k = 100$									
	50	0.063	0.148	0.150	0.152	0.023	0.103	0.107	0.108
	100	0.064	0.153	0.154	0.155	0.024	0.108	0.110	0.111
250	150	0.067	0.158	0.158	0.159	0.026	0.112	0.114	0.115
	200	0.075	0.168	0.168	0.170	0.030	0.121	0.123	0.123
	av	0.062	0.150	0.151	0.152	0.023	0.103	0.106	0.106
$k = 250$									
	100	0.034	0.092	0.093	0.093	0.009	0.063	0.064	0.064
	200	0.035	0.094	0.094	0.095	0.009	0.065	0.065	0.066
500	300	0.036	0.096	0.096	0.096	0.010	0.066	0.067	0.067
	400	0.039	0.100	0.100	0.100	0.011	0.070	0.070	0.070
	av	0.034	0.092	0.093	0.093	0.009	0.062	0.063	0.063

Table 6: Average ratio $\widetilde{c}_{n-m}/\widehat{c}_n$ of condition numbers, uniform design.

		Gaussian			t(5)		
		α			α		
n	m	0	0.05	0.1	0	0.05	0.1
$k = 30$							
	10	0.028	0.027	0.026	0.014	0.013	0.013
	20	0.030	0.029	0.029	0.015	0.014	0.014
50	30	0.034	0.033	0.034	0.017	0.017	0.017
	40	0.045	0.044	0.046	0.023	0.023	0.023
	av	0.028	0.027	0.029	0.014	0.013	0.013
$k = 100$							
	50	0.063	0.069	0.087	0.023	0.023	0.028
	100	0.064	0.092	0.109	0.024	0.027	0.040
250	150	0.067	0.105	0.120	0.026	0.031	0.051
	200	0.075	0.116	0.129	0.030	0.039	0.062
	av	0.062	0.091	0.106	0.023	0.027	0.042
$k = 250$							
	100	0.034	0.058	0.025	0.009	0.012	0.009
	200	0.035	0.070	0.032	0.009	0.018	0.015
500	300	0.036	0.076	0.041	0.010	0.024	0.022
	400	0.039	0.081	0.042	0.011	0.030	0.024
	av	0.034	0.069	0.031	0.009	0.019	0.015

Table 7: Comparison of alternative estimators' precision $R_{2,S}$ for $n = 100$, Toeplitz design.

		Gaussian				t(5)			
		ρ				ρ			
k	Estimator	0	0.5	0.75	0.95	0	0.5	0.75	0.95
30	Sample variance	4.91	5.01	5.29	6.89	23.1	21.9	20.9	24.7
	Our estimator	0.07	3.51	5.13	7.35	1.04	5.76	9.79	20.4
	Shrinkage identity	0.04	3.38	4.76	7.01	1.19	6.16	9.93	16.4
	Shrinkage equicorrelation	0.62	3.53	4.75	6.96	3.43	7.67	10.9	21.6
	Hard thresholding	0.68	4.96	6.19	6.89	4.04	12.9	16.0	20.9
	Soft thresholding	0.59	3.89	5.50	7.06	3.38	10.4	14.3	19.6
100	Sample variance	51.0	51.3	52.3	59.3	137	138	144	181
	Our estimator	0.13	21.2	38.3	64.3	1.83	25.4	53.7	137
	Shrinkage identity	0.07	20.3	37.6	57.0	1.77	27.1	61.7	123
	Shrinkage equicorrelation	2.03	21.3	37.3	56.5	7.02	31.1	63.2	134
	Hard thresholding	2.06	30.0	42.3	64.0	7.55	40.1	117	164
	Soft thresholding	1.99	20.1	35.3	66.8	6.87	38.5	84.2	165
250	Sample variance	313	315	317	335	835	800	801	847
	Our estimator	0.24	68.3	168	301	12.6	76.6	197	514
	Shrinkage identity	0.19	66.1	161	295	10.7	79.4	222	588
	Shrinkage equicorrelation	5.13	69.6	160	291	32.3	93.0	229	575
	Hard thresholding	5.00	86.0	144	378	32.7	104	333	954
	Soft thresholding	4.97	63.0	124	314	31.1	103	272	803

Table 8: Comparison of alternative estimators' precision $\mathrm{R}_{2,\mathrm{S}}$ for $n = 100$, equicorrelation design.

		Gaussian				t(5)			
		ρ				ρ			
k	Estimator	0	0.5	0.75	0.95	0	0.5	0.75	0.95
30	Sample variance	4.91	5.76	7.14	9.24	23.1	15.3	22.3	31.3
	Our estimator	0.07	4.81	7.06	9.30	1.04	11.8	20.6	29.8
	Shrinkage identity	0.04	5.79	7.22	9.21	1.19	12.6	17.3	21.3
	Shrinkage equicorrelation	0.62	3.46	6.36	9.20	3.43	9.38	20.3	31.2
	Hard thresholding	0.68	5.76	7.14	9.24	4.04	15.2	22.0	31.0
	Soft thresholding	0.59	5.86	7.18	9.33	3.38	16.9	22.3	26.3
100	Sample variance	51.0	61.8	79.3	89.8	137	200	551	360
	Our estimator	0.13	52.0	77.7	90.6	1.83	137	421	372
	Shrinkage identity	0.07	63.0	79.2	90.2	1.77	137	180	224
	Shrinkage equicorrelation	2.03	34.9	70.1	89.3	7.02	112	507	359
	Hard thresholding	2.06	61.8	79.3	89.8	7.55	174	236	298
	Soft thresholding	1.99	62.8	79.7	90.8	6.87	182	240	304
250	Sample variance	313	389	495	577	835	1256	1653	1734
	Our estimator	0.24	320	481	574	12.6	794	1487	1647
	Shrinkage identity	0.19	389	494	572	10.7	853	1124	1346
	Shrinkage equicorrelation	5.13	218	437	574	32.3	618	1480	1725
	Hard thresholding	5.00	389	495	577	32.7	1060	1544	1734
	Soft thresholding	4.97	392	499	582	31.1	1069	1459	1835

Table 9: Comparison of alternative estimators' precision $\mathrm{R}_{2,\mathrm{S}}$ for $n = 100$, uniform design.

		Gaussian			t(5)		
		α			α		
k	Estimator	0	0.05	0.1	0	0.05	0.1
30	Sample variance	4.91	4.91	4.92	23.1	22.7	22.2
	Our estimator	0.07	0.39	0.99	1.04	1.61	2.43
	Shrinkage identity	0.04	0.36	1.13	1.19	1.46	2.32
	Shrinkage equicorrelation	0.62	0.72	1.00	3.43	3.42	3.72
	Hard thresholding	0.68	1.03	2.06	4.04	4.18	5.20
	Soft thresholding	0.59	0.94	1.98	3.38	3.57	4.56
100	Sample variance	51.0	51.0	51.1	137	125	125
	Our estimator	0.13	3.37	9.11	1.83	5.55	15.2
	Shrinkage identity	0.07	3.91	12.7	1.77	5.52	15.9
	Shrinkage equicorrelation	2.03	3.20	6.35	7.02	7.64	11.4
	Hard thresholding	2.06	6.21	18.7	7.55	10.82	23.2
	Soft thresholding	1.99	6.13	18.6	6.87	10.31	22.7
250	Sample variance	313	313	314	835	910	909
	Our estimator	0.24	18.7	53.6	12.6	29.4	90.3
	Shrinkage identity	0.19	24.3	79.2	10.7	30.0	95.4
	Shrinkage equicorrelation	5.13	12.2	31.6	32.3	27.8	51.0
	Hard thresholding	5.00	31.1	109	32.7	45.9	124
	Soft thresholding	4.97	31.0	109	31.1	44.9	123

Table 10: Comparison of alternative estimators' condition number for $n = 100$, Toeplitz design.

		Gaussian				t(5)			
		ρ				ρ			
k	Estimator	0	0.5	0.75	0.95	0	0.5	0.75	0.95
30	Sample variance	10.1	32.6	135	2082	23.6	56.8	203	2804
	Our estimator	1.23	4.62	23.0	416	1.32	4.27	21.3	403
	Shrinkage identity	1.08	7.07	37.6	445	1.09	5.19	23.8	258
	Shrinkage equicorrelation	1.85	8.24	38.0	316	2.45	7.34	28.5	216
100	Sample variance	-	-	-	-	-	-	-	-
	Our estimator	1.20	3.22	13.5	341	1.30	2.98	12.4	329
	Shrinkage identity	1.08	4.86	25.1	481	1.20	4.30	16.8	254
	Shrinkage equicorrelation	2.08	7.03	31.8	468	2.97	7.35	25.6	302
250	Sample variance	-	-	-	-	-	-	-	-
	Our estimator	1.19	2.59	8.79	198	1.30	2.45	7.97	187
	Shrinkage identity	1.13	3.47	15.0	291	1.50	4.11	13.1	164
	Shrinkage equicorrelation	2.28	6.09	23.5	385	3.54	7.70	22.9	260

Table 11: Comparison of alternative estimators' condition number for $n = 100$, equicorrelation design.

		Gaussian				t(5)			
		ρ				ρ			
k	Estimator	0	0.5	0.75	0.95	0	0.5	0.75	0.95
30	Sample variance	10.1	135	394	2497	23.6	186	547	3422
	Our estimator	1.23	32.6	97.5	622	1.32	31.9	97.5	621
	Shrinkage identity	1.08	92.9	263	874	1.09	80.0	220	572
	Shrinkage equicorrelation	1.85	40.5	112	635	2.45	44.9	123	669
100	Sample variance	-	-	-	-	-	-	-	-
	Our estimator	1.20	104	319	2025	1.30	102	317	2020
	Shrinkage identity	1.08	995	2786	4708	1.20	509	1444	2570
	Shrinkage equicorrelation	2.08	140	390	2148	2.97	161	449	2295
250	Sample variance	-	-	-	-	-	-	-	-
	Our estimator	1.19	261	795	5083	1.30	254	792	5055
	Shrinkage identity	1.13	2518	6912	11756	1.50	1271	3565	6524
	Shrinkage equicorrelation	2.28	368	1008	5484	3.54	418	1137	5795

Table 12: Comparison of alternative estimators' condition number for $n = 100$, uniform design.

		Gaussian			t(5)		
		α			α		
k	Estimator	0	0.05	0.1	0	0.05	0.1
30	Sample variance	10.1	10.9	14.0	23.6	25.0	27.9
	Our estimator	1.23	1.32	1.93	1.32	1.36	1.63
	Shrinkage identity	1.08	1.21	1.82	1.09	1.18	1.57
	Shrinkage equicorrelation	1.85	2.48	3.55	2.45	2.99	4.09
100	Sample variance	-	-	-	-	-	-
	Our estimator	1.20	2.03	4.63	1.30	1.47	2.72
	Shrinkage identity	1.08	1.52	3.57	1.20	1.54	2.83
	Shrinkage equicorrelation	2.08	5.34	9.54	2.97	6.10	10.7
250	Sample variance	-	-	-	-	-	-
	Our estimator	1.19	4.06	10.5	1.30	1.88	5.07
	Shrinkage identity	1.13	2.05	6.75	1.50	2.32	5.19
	Shrinkage equicorrelation	2.28	11.5	22.7	3.54	13.2	25.3

Table 13: Bootstrap-based choice of m, for $k = 30, n = 50$, Gaussian distribution.[a]

	$\rho = 0$				$\rho = 0.5$				$\rho = 0.75$				$\rho = 0.95$			
$(m/n)^d$	d =0	0.1	0.25	0.5	0	0.1	0.25	0.5	0	0.1	0.25	0.5	0	0.1	0.25	0.5
Mean	27.3	23.9	7.0	5.0	26.6	23.7	18.6	5.9	23.3	20.8	16.4	6.7	26.5	25.9	24.6	16.6
Std. dev.	0.79	0.89	3.08	0.00	0.92	0.84	0.94	0.33	1.51	1.50	1.43	1.15	3.37	3.61	4.59	8.99
Median	27	24	6	5	27	24	19	6	23	21	16	6	27	26	25	20
Minimum	24	21	5	5	24	21	14	5	20	17	13	6	11	11	5	5
10% quantile	26	23	6	5	25	23	18	5	22	19	14.9	6	22	21	19	5
25% quantile	27	23	6	5	26	23	18	6	22	20	15	6	24	24	22	5
70% quantile	28	24	6	5	27	24	19	6	24	22	17	7	29	28	28	24
90% quantile	28	25	7	5	28	25	19	6	25	23	18	7	31	30	30	27
Maximum	30	27	19	5	29	27	21	6	29	27	24	15	36	36	36	36
MSE increase	127%	91.2%	3.4%	0.0%	0.1%	0.2%	2.1%	29.7%	0.0%	0.2%	3.5%	52.3%	2.4%	0.0%	1.2%	2.4%

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 $\mathrm{R}_{2,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}}_{m,S})$. 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 $\rho = 0.5$ and $d = 0$, the bootstrap suggests taking $m_B = 27$ which results in the MSE being 5.390, while the optimal MSE at $m = 26$ is 5.383, hence an increase of 0.1%. All results are based on 1,000 simulations and 1,000 bootstrap replications.

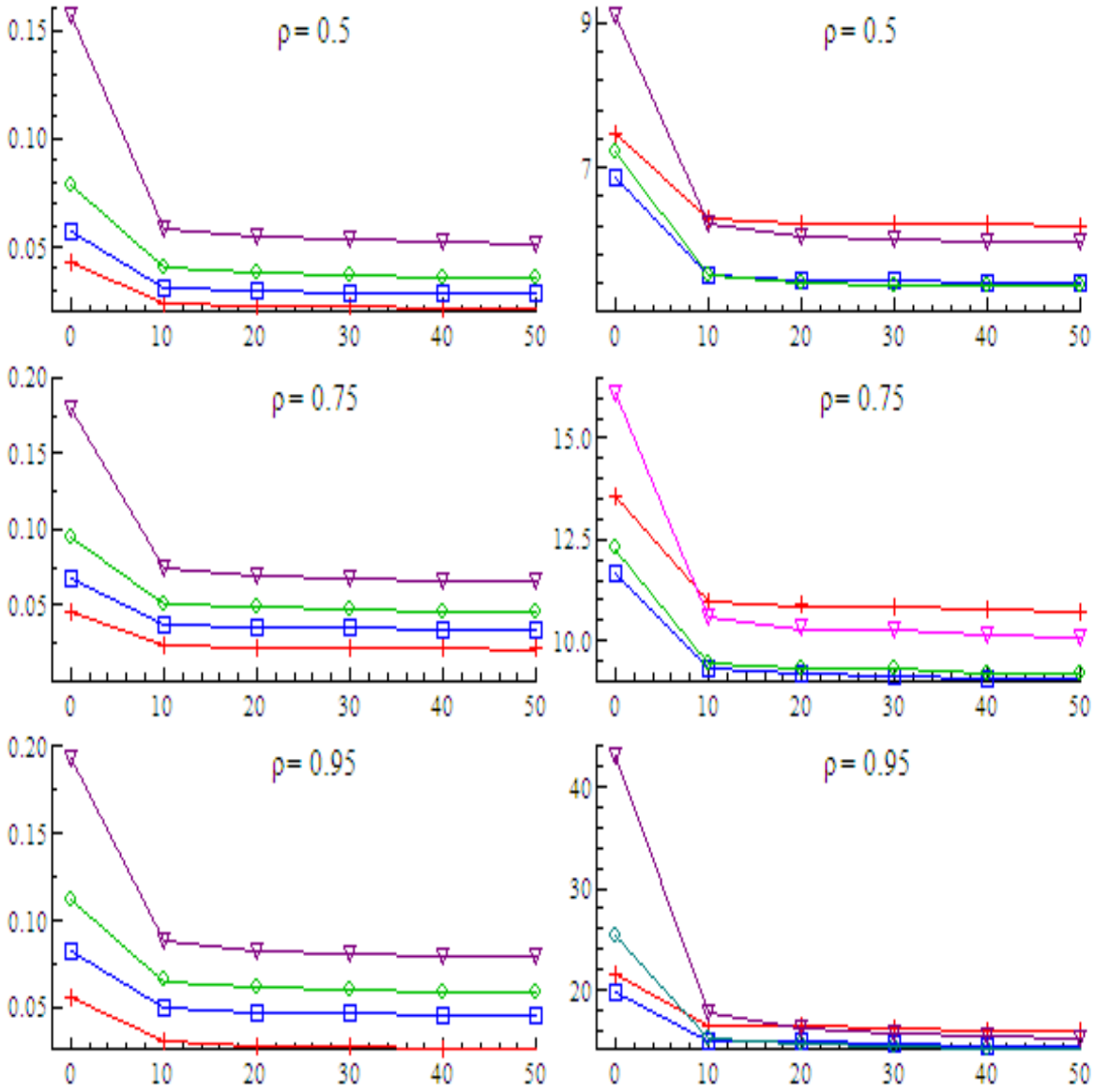

Figure 1: Ratio of condition numbers $\widetilde{c}_{n-m}/\widehat{c}_n$ (left panel) and $\mathrm{R}_{2,\mathrm{S}}(\widetilde{\boldsymbol{\Sigma}}_{m,S})$ (right panel), averaged over S simulations (horizontal axes), for $k = 30$, $n = 50$, and $\boldsymbol{x} \sim \mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$.