Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2013 June 1; 75(3): 427–450. doi:10.1111/j.1467-9868.2012.01049.x.

Condition Number Regularized Covariance Estimation*

Joong-Ho Won,
School of Industrial Management Engineering, Korea University, Seoul, Korea

Johan Lim,
Department of Statistics, Seoul National University, Seoul, Korea

Seung-Jean Kim, and
Citi Capital Advisors, New York, NY, USA

Bala Rajaratnam
Department of Statistics, Stanford University, Stanford, CA, USA

Abstract

Estimation of high-dimensional covariance matrices is known to be a difficult problem, has many applications, and is of current interest to the larger statistics community. In many applications including so-called the "large p small n" setting, the estimate of the covariance matrix is required to be not only invertible, but also well-conditioned. Although many regularization schemes attempt to do this, none of them address the ill-conditioning problem directly. In this paper, we propose a maximum likelihood approach, with the direct goal of obtaining a well-conditioned estimator. No sparsity assumption on either the covariance matrix or its inverse are are imposed, thus making our procedure more widely applicable. We demonstrate that the proposed regularization scheme is computationally efficient, yields a type of Steinian shrinkage estimator, and has a natural Bayesian interpretation. We investigate the theoretical properties of the regularized covariance estimator comprehensively, including its regularization path, and proceed to develop an approach that adaptively determines the level of regularization that is required. Finally, we demonstrate the performance of the regularized estimator in decision-theoretic comparisons and in the financial portfolio optimization setting. The proposed approach has desirable properties, and can serve as a competitive procedure, especially when the sample size is small and when a well-conditioned estimator is required.

Keywords

covariance estimation; regularization; convex optimization; condition number; eigenvalue; shrinkage; cross-validation; risk comparisons; portfolio optimization

1 Introduction

We consider the problem of regularized covariance estimation in the Gaussian setting. It is well known that, given n independent samples $x_1, \ldots, x_n \in \mathbb{R}^p$ from a zero-mean p-variate Gaussian distribution, the sample covariance matrix

*A preliminary version of the paper has appeared in an unrefereed conference proceedings previously.

Supplemental materials Accompanying supplemental materials contain additional proofs of Theorems 2 and 3, and Proposition 2; additional figures illustrating Bayesian prior densities; additional figure illustrating risk simulations; details of the empirical minimum variance rebalancing study.

$$S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T,$$

maximizes the log-likelihood as given by

$$\begin{aligned} l(\sum) &= \log \prod_{i=1}^{n} \frac{1}{(2\pi)^p (\det \sum)^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2} x_i^T \sum^{-1} x_i\right) \\ &= -(np/2)\log(2\pi) - (n/2)(\mathbf{tr}(\sum^{-1} S) - \log\det\sum^{-1}), \end{aligned} \quad (1)$$

where det A and $\mathbf{tr}(A)$ denote the determinant and trace of a square matrix A respectively. In recent years, the availability of high-throughput data from various applications has pushed this problem to an extreme where in many situations, the number of samples, n, is often much smaller than the dimension of the estimand, p. When $n < p$ the sample covariance matrix S is singular, not positive definite, and hence cannot be inverted to compute the precision matrix (the inverse of the covariance matrix), which is also needed in many applications. Even when $n > p$, the eigenstructure tends to be systematically distorted unless p/n is extremely small, resulting in numerically ill-conditioned estimators for Σ; see Dempster (1972) and Stein (1975). For example, in mean-variance portfolio optimization (Markowitz, 1952), an ill-conditioned covariance matrix may amplify estimation error because the optimal portfolio involves matrix inversion (Ledoit and Wolf, 2003; Michaud, 1989). A common approach to mitigate the problem of numerical stability is regularization.

In this paper, we propose regularizing the sample covariance matrix by explicitly imposing a constraint on the condition number.[1] Instead of using the standard estimator S, we propose to solve the following constrained maximum likelihood (ML) estimation problem

$$\begin{aligned} &\text{maximize} && l(\sum) \\ &\text{subject to} && \text{cond}(\sum) \le \kappa_{\max}, \end{aligned} \quad (2)$$

where cond(M) stands for the condition number, a measure of numerical stability, of a matrix M (see Section 1.1 for details). The matrix M is invertible if cond(M) is finite, ill-conditioned if cond(M) is finite but high (say, greater than 10^3 as a rule of thumb), and well-conditioned if cond(M) is moderate. By bounding the condition number of the estimate by a regularization parameter $\kappa_{\max}$, we directly address the problem of invertibility or ill-conditioning. This direct control is appealing because the true covariance matrix is in most situations unlikely to be ill-conditioned whereas its sample counterpart is often ill-conditioned. It turns out that the resulting regularized matrix falls into a broad family of Steinian-type shrinkage estimators that shrink the eigenvalues of the sample covariance matrix towards a given structure (James and Stein, 1961; Stein, 1956). Moreover, the regularization parameter $\kappa_{\max}$ is adaptively selected from the data using cross validation.

Numerous authors have explored alternative estimators for Σ (or Σ^{-1}) that perform better than the sample covariance matrix S from a decision-theoretic point of view. Many of these estimators give substantial risk reductions compared to S in small sample sizes, and often involve modifying the spectrum of the sample covariance matrix. A simple example is the family of linear shrinkage estimators which take a convex combination of the sample

[1]This procedure was first considered by two of the authors of this paper in a previous conference paper and is further elaborated in this paper (see Won and Kim (2006)).

covariance matrix and a suitably chosen target or regularization matrix. Notable in the area is the seminal work of Ledoit and Wolf (2004) who study a linear shrinkage estimator toward a specified target covariance matrix, and choose the optimal shrinkage to minimize the Frobenius risk. Bayesian approaches often directly yield estimators which shrink toward a structure associated with a pre-specified prior. Standard Bayesian covariance estimators yield a posterior mean Σ that is a linear combination of S and the prior mean. It is easy to show that the eigenvalues of such estimators are also linear shrinkage estimators of Σ; see, *e.g.*, Haff (1991). Other nonlinear Steinian-type estimators have also been proposed in the literature. James and Stein (1961) study a constant risk minimax estimator and its modification in a class of orthogonally invariant estimators. Dey and Srinivasan (1985) provide another minimax estimator which dominates the James-Stein estimator. Yang and Berger (1994) and Daniels and Kass (2001) consider a reference prior and hierarchical priors, that respectively yield posterior shrinkage.

Likelihood-based approaches using multivariate Gaussian models have provided different perspectives on the regularization problem. Warton (2008) derives a novel family of linear shrinkage estimators from a penalized maximum likelihood framework. This formulation enables cross-validation of the regularization parameter, which we discuss in Section 3 for the proposed method. Related work in the area include Sheena and Gupta (2003), Pourahmadi et al. (2007), and Ledoit and Wolf (2012). An extensive literature review is not undertaken here, but we note that the approaches mentioned above (and the one proposed in this paper) fall in the class of covariance estimation and related problems which do not assume or impose sparsity, on either the covariance matrix, or its inverse (for such approaches either in the frequentist, Bayesian, or testing frameworks, the reader is referred to Banerjee et al. (2008); Friedman et al. (2008); Hero and Rajaratnam (2011, 2012); Khare and Rajaratnam (2011); Letac and Massam (2007); Peng et al. (2009); Rajaratnam et al. (2008)).

1.1 Regularization by shrinking sample eigenvalues

We briefly review Steinian-type eigenvalue shrinkage estimators in this subsection. Dempster (1972) and Stein (1975) noted that the eigenstructure of the sample covariance matrix S tends to be systematically distorted unless p/n is extremely small. They observed that the larger eigenvalues of S are overestimated whereas the smaller ones are underestimated. This observation led to estimators which directly modify the spectrum of the sample covariance matrix and are designed to "shrink" the eigenvalues together. Let l_i, $i = 1, \ldots, p$, denote the eigenvalues of the sample covariance matrix (sample eigenvalues) in nonincreasing order ($l_1 \geq \ldots \geq l_p \geq 0$). The spectral decomposition of the sample covariance matrix is given by

$$S = Q \operatorname{diag}(l_1, \ldots, l_p)\, Q^T, \quad (3)$$

where $\operatorname{diag}(l_1, \ldots, l_p)$ is the diagonal matrix with diagonal entries l_i and $Q \in \mathbb{R}^{p \times p}$ is the orthogonal matrix whose i-th column is the eigenvector that corresponds to the eigenvalue l_i. As discussed above, a large number of covariance estimators regularizes S by modifying its eigenvalues with the explicit goal of better estimating the eigenspectrum. In this light Stein (1975) proposed the class of orthogonally invariant estimators of the following form:

$$\widehat{\sum} = Q \operatorname{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p) Q^T. \quad (4)$$

Typically, these estimators shrink the sample eigenvalues so that the modified eigenspectrum is less spread than that of the sample covariance matrix. In many estimators, the shrunk eigenvalues are required to maintain the original order as those of S: $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$.

One well-known example of Steinian-type shrinkage estimators is the linear shrinkage estimator as given by

$$\widehat{\sum}_{\mathrm{LS}} = (1-\delta)S + \delta F, \quad 0 \leq \delta \leq 1 \quad (5)$$

where the target matrix $F = cI$ for some $c > 0$ (Ledoit and Wolf, 2004; Warton, 2008). For the linear estimator the relationship between the sample eigenvalues l_i and the modified eigenvalues $\hat{\lambda}_i$ is affine:

$$\widehat{\lambda}_i = (1-\delta)l_i + \delta c$$

Another example, Stein's estimator (Stein, 1975, 1986), denoted by $\hat{\Sigma}_{\mathrm{Stein}}$, is given by $\hat{\lambda}_i = l_i/d_i$, $i = 1, \ldots, p$, with $d_i = (n - p + 1 + 2l_i \Sigma_{j \neq i}(l_i - l_j)^{-1})/n$. The original order in the estimator is preserved by applying isotonic regression (Lin and Perlman, 1985).

1.2 Regularization by imposing a condition number constraint

Now we proceed to introduce the condition number-regularized covariance estimator proposed in this paper. Recall that the condition number of a positive definite matrix Σ is defined as

$$\mathrm{cond}(\sum) = \lambda_{\max}(\sum)/\lambda_{\min}(\sum)$$

where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are the maximum and the minimum eigenvalues of Σ, respectively. (Understand that $\mathrm{cond}(\Sigma) = \infty$ if $\lambda_{\min}(\Sigma) = 0$.) The condition number-regularized covariance estimation problem (2) can therefore be formulated as

$$\begin{array}{ll} \text{maximize} & l(\sum) \\ \text{subject to} & \lambda_{\max}(\sum)/\lambda_{\min}(\sum) \leq \kappa_{\max}. \end{array} \quad (6)$$

An implicit condition is that Σ be symmetric and positive definite[2].

The estimation problem (6) can be reformulated as a convex optimization problem in the matrix variable Σ^{-1} (see Section 2). Standard methods such as interior-point methods can solve the convex problem when the number of variables (i.e., entries in the matrix) is modest, say, under 1000. Since the number of variables is about $p(p + 1)/2$, the limit is around $p = 45$. Such a naive approach is not adequate for moderate to high dimensional problems. One of the main contributions of this paper is a significant improvement on the solution method for (6) so that it scales well to much larger sizes. In particular, we show that

[2]This problem can be considered a generalization of the problem studied by Sheena and Gupta (2003), who consider imposing either a fixed lower bound *or* a fixed upper bound on the eigenvalues. Their approach is however different from ours in a fundamental sense in that it is not designed to control the condition number. Hence such a solution does not correct for the overestimation of the largest eigenvalues and underestimation of the small eigenvalues simultaneously.

(6) reduces to an unconstrained univariate optimization problem. Furthermore, the solution to (6), denoted by $\hat{\Sigma}_{\text{cond}}$, has a Steinian-type shrinkage of the form as in (4) with eigenvalues given by

$$\widehat{\lambda_i} = \min\left(\max(\tau^\star, l_i), \kappa_{\max}\tau^\star\right) = \begin{cases} \tau^\star, & l_i \le \tau^\star \\ l_i, & \tau^\star < l_i < \kappa_{\max}\tau^\star \\ \kappa_{\max}\tau^\star, & l_i \ge \kappa_{\max}\tau^\star, \end{cases} \quad (7)$$

for some $\tau^\star > 0$. Note that the modified eigenvalues are nonlinear functions of the sample eigenvalues, meet the order constraint of Section 1.1, and even when $n < p$, the nonlinear shrinkage estimator $\hat{\Sigma}_{\text{cond}}$ is well-conditioned. The quantity $\tau^\star$ is determined adaptively from the data and the choice of the regularization parameter $\kappa_{\max}$. This solution method was first considered by two of the authors of this paper in a previous conference proceeding (Won and Kim, 2006). In this paper, we give a formal proof of the assertion that the matrix optimization problem (6) reduces to an equivalent unconstrained univariate minimization problem. We further elaborate on the proposed method by showing rigorously that $\tau^\star$ can be found exactly and easily with computational effort of order $O(p)$ (Section 2.1).

The nonlinear shrinkage in (7) has a simple interpretation: the eigenvalues of the estimator $\hat{\Sigma}_{\text{cond}}$ are obtained by truncating the sample eigenvalues larger than $\kappa_{\max}\tau^\star$ to $\kappa_{\max}\tau^\star$ and those smaller than $\tau^\star$ to $\tau^\star$. Figure 1 illustrates the functional form of (7) in comparison with that of the linear shrinkage estimator. This novel shrinkage form is rather surprising, because the original motivation of the regularization in (6) is to numerically stabilize the covariance estimate. Note that this type of shrinkage better preserves the eccentricity of the sample estimate than the linear shrinkage which shrinks it toward a spherical covariance matrix.

Other major contributions of the paper include a detailed analysis of the regularization path of the shrinkage enabled by a geometric perspective of the condition number-regularized estimator (Section 2.2), and an adaptive selection procedure for choosing the regularization parameter $\kappa_{\max}$ under the maximum predictive likelihood criterion (Section 3) and properties thereof.

We also undertake a Bayesian analysis of the condition number constrained estimator (Section 4). A detailed application of the methodology to real data, which demonstrates the usefulness of the method in a practical setting, is also provided (Section 6). In particular, Section 6 studies the use of the proposed estimator in the mean-variance portfolio optimization setting. Section 5 undertakes a simulation study in order to compare the risk of the condition number-regularized estimator to those of others. Risk comparisons in higher dimensional settings (as compared to those given in the preliminary conference paper) are provided in this Section. Asymptotic properties of the estimator are also investigated in Section 5.

2 Condition number-regularized covariance estimation

2.1 Derivation and characterization of solution

This section gives the derivation of the solution (7) of the condition number-regularized covariance estimation problem as given in (6) and shows how to compute $\tau^\star$ given $\kappa_{\max}$. Note that it suffices to consider the case $\kappa_{\max} < l_1/l_p = \text{cond}(S)$, since otherwise, the condition number constraint is already satisfied and the solution to (6) simply reduces to the sample covariance matrix S.

It is well known that the log-likelihood (1) of a multivariate Gaussian covariance matrix is a convex function of $\Omega = \Sigma^{-1}$. Note that Ω is the canonical parameter for the (Wishart) natural exponential family associated with the likelihood in (1). Since $\mathrm{cond}(\Sigma) = \mathrm{cond}(\Omega)$, it can be shown that the condition number constraint on Ω is equivalent to the existence of $u > 0$ such that $uI \preceq \Omega \preceq \kappa_{\max} uI$, where $A \preceq B$ denotes that $B - A$ is positive semidefinite (Boyd and Vandenberghe, 2004, Chap. 7). Therefore the covariance estimation problem (6) is equivalent to

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(\Omega S) - \log\det\Omega \\ \text{subject to} & uI \prec \Omega \prec \kappa_{\max} uI, \end{array} \quad (8)$$

where the variables are a symmetric positive definite $p \times p$ matrix Ω and a scalar $u > 0$. The above problem in (8) is a convex optimization problem with $p(p+1)/2 + 1$ variables, *i.e.*, $O(p^2)$.

The following lemma shows that by exploiting structure of the problem it can be reduced to a univariate convex problem, *i.e.*, the dimension of the system is only of $O(1)$ as compared to $O(p^2)$.

Lemma 1—The optimization problem (8) is equivalent to the unconstrained univariate convex optimization problem

$$\mathit{minimize} \sum_{i=1}^{p} J^{(i)}_{\kappa_{\max}}(u), \quad (9)$$

where

$$J^{(i)}_{\kappa_{\max}}(u) = l_i \mu_i^\star(u) - \log \mu_i^\star(u) = \begin{cases} l_i(\kappa_{\max} u) - \log(\kappa_{\max} u), & u < 1/(\kappa_{\max} l_i) \\ 1 + \log l_i, & 1/(\kappa_{\max} l_i) \le u \le 1/l_i \\ l_i u - \log u, & u > 1/l_i, \end{cases}$$

and

$$\mu_i^\star(u) = \min\{\max\{u, 1/l_i\}, \kappa_{\max} u\}, \quad (10)$$

in the sense that the solution to (8) is a function of the solution $u^\star$ to (9), as follows.

$$\Omega^\star = Q \mathrm{diag}(\mu_1^\star(u^\star), \ldots, \mu_p^\star(u^\star)) Q^T,$$

with Q defined as in (3).

Proof: The proof is given in the Appendix.

Characterization of the solution to (9) is given by the following theorem.

Theorem 1—Given $\kappa_{\max} \le \mathrm{cond}(S)$, the optimization problem (9) has a unique solution given by

$$u^\star = \frac{\alpha^\star + p - \beta^\star + 1}{\sum_{i=1}^{\alpha^\star} l_i + \sum_{i=\beta^\star}^{p} \kappa_{\max} l_i}, \quad (11)$$

where $\alpha^\star \in \{1, \ldots, p\}$ is the largest index such that $1/l_{\alpha^\star} < u^\star$ and $\beta^\star \in \{1, \ldots, p\}$ is the smallest index such that $1/l_{\beta^\star} > \kappa_{\max} u^\star$. The quantities $\alpha^\star$ and $\beta^\star$ are not determined a priori but can be found in $O(p)$ operations on the sample eigenvalues $l_1 \geq \ldots \geq l_p$. If $\kappa_{\max} > \text{cond}(S)$, the maximizer $u^\star$ is not unique but $\hat{\Sigma}_{\text{cond}} = S$ for all the maximizing values of u.

Proof: The proof is given in the Appendix.

Comparing (10) to (7), it is immediately seen that

$$\tau^\star = 1/(\kappa_{\max} u^\star) = \frac{\sum_{i=1}^{\alpha^\star} l_i/\kappa_{\max} + \sum_{i=\beta^\star}^{p} l_i}{\alpha^\star + p - \beta^\star + 1}. \quad (12)$$

Note that the lower cutoff level $\tau^\star$ is an average of the (scaled and) truncated eigenvalues, in which the eigenvalues above the upper cutoff level $\kappa_{\max}\tau^\star$ are shrunk by a factor of $1/\kappa_{\max}$.

We highlight the fact that a reformulation of the original minimization into a univariate optimization problem makes the proposed methodology very attractive in high dimensional settings. The method is only limited by the complexity of spectral decomposition of the sample covariance matrix (or the singular value decomposition of the data matrix). The approach proposed here is therefore much faster than using interior point methods. We also note that the condition number-regularized covariance estimator is orthogonally invariant: if $\hat{\Sigma}_{\text{cond}}$ is the estimator of the true covariance matrix Σ, then $U\hat{\Sigma}_{\text{cond}}U^T$ is the estimator of the true covariance matrix $U\Sigma U^T$, for U orthogonal.

2.2 A geometric perspective on the regularization path

In this section, we shall show that a simple relaxation of the optimization problem (8) provides an intuitive geometric perspective on the condition number-regularized estimator. Consider the function

$$J(u, v) = \min_{uI \preceq \Omega \preceq vI} (\mathbf{tr}(\Omega S) - \log\det\Omega) \quad (13)$$

defined as the minimum of the objective of (8) over a range $uI \preceq \Omega \preceq vI$, where $0 < u \leq v$. Note that the relaxation in (13) above differs from the original problem in the sense that the optimization is no longer with respect to the variable u. In particular, u and v are fixed in (13). By fixing u, the problem has now been significantly simplified. Paralleling Lemma 1, it is easily shown that

$$J(u, v) = \sum_{i=1}^{p} \min_{u \leq \mu_i \leq v} (l_i\mu_i - \log\mu_i).$$

Recall that the truncation range of the sample eigenvalues is therefore given by $(1/v, 1/u)$. Now for given u, let $\alpha \in \{1, \ldots, p\}$ be the largest index such that $l_\alpha > 1/u$ and $\beta \in \{1, \ldots, p\}$ be the smallest index such that $l_\beta < 1/v$, *i.e.*, the set of indexes where truncation at either end

of the spectrum first starts to become a binding constraint. With this convention it is easily shown that $J(u, v)$ can now be expressed in simpler form:

$$J(u,v)=\sum_{i=1}^{p}(l_i\mu_i^\star(u,v)-\log\mu_i^\star(u,v)) \quad (14)$$

$$=\sum_{i=1}^{\alpha}(l_iu-\log u)+\sum_{i=\alpha+1}^{\beta-1}(1+\log l_i)+\sum_{i=\beta}^{p}(l_iv-\log v), \quad (15)$$

where

$$\mu_i^\star(u,v)=\min\{\max\{u,1/l_i\},v\}=\begin{cases}u, & 1\le i\le\alpha\\ 1/l_i, & \alpha<i<\beta\\ v, & \beta\le i\le p.\end{cases} \quad (16)$$

Comparing (16) to (10), we observe that $(\Omega^*)^{-1}$, the covariance estimate whose inverse achieves the minimum in (13), is obtained by truncating the eigenvalues of S greater than $1/u$ to $1/u$, and less than $1/v$ to $1/v$.

Furthermore, note that the function $J(u, v)$ has the following properties:

1. J does not increase as u decreases and v increases simultaneously. This follows from noting that simultaneously decreasing u and increasing v expands the domain of the minimization in (13).
2. $J(u, v) = J(1/l_1, 1/l_p)$ for $u \le 1/l_1$ and $v \ge 1/l_p$. Hence $J(u, v)$ is constant on this part of the domain. For these values of u and v, $(\Omega^\star)^{-1} = S$.
3. $J(u, v)$ is almost everywhere differentiable in the interior of the domain $\{(u, v) : 0 < u \le v\}$, except for on the lines $u = 1/l_1,\ldots, 1/l_p$ and $v = 1/l_1,\ldots, 1/l_p$. This follows from noting the the indexes α and β changes their values only on these lines. Hence the contribution to the three summands in (14) changes at these values.

We can now see the following obvious relation between the function $J(u, v)$ and the original problem (8): the solution to (8) as given by $u^\star$ is the minimizer of $J(u, v)$ on the line $v = \kappa_{\max}u$, *i.e.*, the original univariate optimization problem (9) is equivalent to minimizing $J(u, \kappa_{\max}u)$. We denote this minimizer by $u^\star(\kappa_{\max})$ and investigate how $u^\star(\kappa_{\max})$ behaves as $\kappa_{\max}$ varies. The following proposition proves that $u^\star(\kappa_{\max})$ is monotone in $\kappa_{\max}$. This result sheds light on the regularization path, *i.e.*, the solution path of $u^\star$ as $\kappa_{\max}$ varies.

Proposition 1— $u^\star(\kappa_{\max})$ is nonincreasing in $\kappa_{\max}$ and $v^\star(\kappa_{\max}), \triangleq \kappa_{\max}u^\star(\kappa_{\max})$ is nondecreasing, both almost surely.

Proof: The proof is given in the Appendix.

Remark: The above proposition has a very natural and intuitive interpretation: when the constraint on the condition number is relaxed to allow higher value of $\kappa_{\max}$, the gap between $u^\star$ and $v^\star$ widens so that the ratio of $v^\star$ to $u^\star$ remains at $\kappa_{\max}$. This implies that as $\kappa_{\max}$ is increased, the lower truncation value $u^\star$ decreases and the higher truncation value $v^\star$ increases. Proposition 1 can be equivalently interpreted by noting that the optimal truncation range $\tau^\star(\kappa_{\max}), \kappa_{\max}\tau^\star(\kappa_{\max})$ of the sample eigenvalues is nested.

In light of the solution to the condition number-regularized covariance estimation problems in (7), Proposition 1 also implies that once an eigenvalue l_i is truncated for $\kappa_{\max} = \nu_0$, then it remains truncated for all $\kappa_{\max} < \nu_0$. Hence the regularized eigenvalue estimates are not only continuous, but they are also monotonic in the sense that they approach either end of the truncation spectrum as the regularization parameter $\kappa_{\max}$ is decreased to 1. This monotonicity of the condition number-regularized estimates gives a desirable property that is not always enjoyed by other regularized estimators, such as the lasso for instance (Personal communications: Jerome Friedman, Department of Statistics, Stanford University).

With the above theoretical knowledge on the regularization path of the sample eigenvalues, we now provide an illustration of the regularization path; see Figure 2. More specifically, consider the plot of the path of the optimal point $(u^\star(\kappa_{\max}), v^\star(\kappa_{\max}))$ on the u–v plane from $(u^\star(1), u^\star(1))$ to $(1/l_1, 1/l_p)$ when varying $\kappa_{\max}$ from 1 to cond(S). The left panel shows the path of $(u^\star(\kappa_{\max}), v^\star(\kappa_{\max}))$ on the u–v plane for the case where the sample eigenvalues are (21, 7, 5.25, 3.5, 3). Here a point on the path represents the minimizer of $J(u, v)$ on a line $v = \kappa_{\max} u$ (hollow circle). The path starts from a point on the solid line $v = u$ ($\kappa_{\max} = 1$, square) and ends at $(1/l_1, 1/l_p)$, where the dashed line $v = \text{cond}(S) \cdot u$ passes ($\kappa_{\max} = \text{cond}(S)$, solid circle). Note that the starting point ($\kappa_{\max} = 1$) corresponds to $\hat{\Sigma}_{\text{cond}} = \gamma I$ for some data-dependent $\gamma > 0$ and the end point ($\kappa_{\max} = \text{cond}(S)$) to $\hat{\Sigma}_{\text{cond}} = S$. When $\kappa_{\max} > \text{cond}(S)$, multiple values of $u^\star$ are achieved in the shaded region above the dashed line, yielding nevertheless the same estimator S. The right panel of Figure 2 shows how the eigenvalues of the estimated covariance vary as a function of $\kappa_{\max}$. Here we see that as the constraint is made stricter the eigenvalue estimates decrease monotonically. Furthermore, the truncation range of the eigenvalues simultaneously decreases and remains nested.

3 Selection of regularization parameter $\kappa_{\max}$

Sections 2.1 and 2.2 have already discussed how the optimal truncation range $(\tau^\star, \kappa_{\max}\tau^\star)$ is determined for a given regularization parameter $\kappa_{\max}$, and how it varies with the value of $\kappa_{\max}$. This section proposes a data-driven (or adaptive) criterion for selecting the optimal $\kappa_{\max}$ and undertakes an analysis of this approach.

3.1 Predictive risk selection procedure

A natural approach is to select $\kappa_{\max}$ that minimizes the *predictive risk*, or the expected negative predictive likelihood as given by

$$\text{PR}(\nu) = \mathbf{E}\left[\mathbf{E}_{\tilde{x}}\left\{\mathbf{tr}\left(\sum\nolimits_{\nu}^{-1}\tilde{X}\tilde{X}^T\right) - \text{logdet}\widehat{\sum}\nolimits_{\nu}^{-1}|\mathrm{x}_1, \ldots, \mathrm{x}_n\right\}\right], \quad (17)$$

where $\hat{\Sigma}_\nu$ is the estimated condition number-regularized covariance matrix from independent samples $x_1, \cdots, x_n$, with the value of the regularization parameter $\kappa_{\max}$ set to ν, and $\tilde{X} \in \mathbb{R}^p$ is a random vector, independent of the given samples, from the same distribution. We approximate the predictive risk using K-fold cross validation. The K-fold cross validation approach divides the data matrix $\mathbf{X} = (x_1^T, \cdots, x_n^T)$ into K groups so that $\mathbf{X}^T = (X_1^T, \ldots, X_K^T)$ with n_k observations in the k-th group, $k = 1, \ldots, K$. For the k-th iteration, each observation in the k-th group X_k plays the role of $\tilde{X}$ in (17), and the remaining $K - 1$ groups are used together to estimate the covariance matrix, denoted by $\widehat{\sum}_{\nu}^{[-k]}$. The approximation of the predictive risk using the k-th group reduces to the predictive log-likelihood

$$l_k(\widehat{\sum}_{\nu}^{[-k]}, X_k) = -(n_k/2)\left[\mathbf{tr}\{(\widehat{\sum}_{\nu}^{[-k]})^{-1} X_k X_k^T/n_k\} - \log\det(\widehat{\sum}_{\nu}^{[-k]})^{-1}\right].$$

The estimate of the predictive risk is then defined as

$$\widehat{\mathrm{PR}}(\nu) = -\frac{1}{n}\sum_{k=1}^{K} l_k(\widehat{\sum}_{\nu}^{[-k]}, X_k). \quad (18)$$

The optimal value for the regularization parameter $\kappa_{\max}$ is selected as ν that minimizes (18), *i.e.*,

$$\widehat{\kappa}_{\max} = \inf(\arg\inf_{\nu \geq 1} \widehat{\mathrm{PR}}(\nu)).$$

Note that the outer infimum is taken since $l_k(\widehat{\sum}_{\nu}^{[-k]}, X_k)$ is constant for $\nu \geq \mathrm{cond}(S^{[-k]})$, where $S^{[-k]}$ is the k-th fold sample covariance matrix based on the remaining $K - 1$ groups.

3.2 Properties of the optimal regularization parameter

We proceed to investigate the behavior of the selected regularization parameter $\hat{\kappa}_{\max}$, both theoretically and in simulations. We first note that $\hat{\kappa}_{\max}$ is a consistent estimator for the true condition number κ. This fact is expressed below as one of the several properties of $\hat{\kappa}_{\max}$:

(P1) For a fixed p, as n increases, $\hat{\kappa}_{\max}$ approaches κ in probability, where κ is the condition number of the true covariance matrix Σ.

This statement is stated formally below:

Theorem 2—For a given p,

$$\lim_{n\to\infty} \mathbf{pr}(\widehat{\kappa}_{\max} = \kappa) = 1.$$

Proof: The proof is given in Supplemental Section A.

We further investigate the behavior of $\kappa_{\max}$ in simulations. To this end, consider *iid* samples from a zero-mean p-variate Gaussian distribution with the following covariance matrices:

- **i.** Identity matrix in $\mathbb{R}^p$.
- **ii.** diag(1, r, r^2, …, r^p), with condition number $1/r^p = 5$.
- **iii.** diag(1, r, r^2, …, r^p), with condition number $1/r^p = 400$.
- **iv.** Toeplitz matrix whose (i, j)th element is $0.3^{|i-j|}$ for $i, j = 1, \ldots, p$.

We consider different combinations of sample sizes and dimensions of the problem as given by $n = 20, 80, 320$ and $p = 5, 20, 80$. For each of these cases 100 replicates are generated and $\hat{\kappa}_{\max}$ computed with 5-fold cross validation. The behavior of $\hat{\kappa}_{\max}$ is plotted in Figure 3 and

leads to insightful observations. A summary of the properties of the $\hat{\kappa}_{\max}$ is given in (P2)–(P4) below:

- **(P2)** If the condition number of the true covariance matrix remains finite as p increases, then for a fixed n, $\hat{\kappa}_{\max}$ decreases.
- **(P3)** If the condition number of the true covariance matrix remains finite as p increases, then for a fixed n, $\hat{\kappa}_{\max}$ converges to 1.
- **(P4)** The variance of $\hat{\kappa}_{\max}$ decreases as either n or p increases.

These properties are analogous to those of the optimal regularization parameter $\hat{\delta}$ for the linear shrinkage estimator (5), found using a similar predictive risk criterion (Warton, 2008).

4 Bayesian interpretation

In the same spirit as the Bayesian posterior mode interpretation of the LASSO (Tibshirani, 1996), we can draw parallels for the condition number regularized covariance estimator. The condition number constraint given by $\lambda_1(\Sigma)/\lambda_p(\Sigma) \leq \kappa_{\max}$ is similar to adding a penalty term $g_{\max}(\lambda_1(\Sigma)/\lambda_p(\Sigma))$ to the likelihood equation for the eigenvalues:

$$\begin{array}{ll} \text{maximize} & \exp\left(-(n/2)\sum_{i=1}^{p} l_i/\lambda_i\right)\left(\prod_{i=1}^{p}\lambda_i\right)^{-n/2}\exp(-g_{\max}\,\lambda_1/\lambda_p) \\ \text{subject to} & \lambda_1 \geq \cdots \geq \lambda_p > 0 \end{array}$$

The above expression allows us to qualitatively interpret the condition number-regularized estimator as the Bayes posterior mode under the following prior

$$\pi(\lambda_1,\ldots,\lambda_p)=\exp(-g_{\max}\,\lambda_1/\lambda_p), \quad \lambda_1 \geq \cdots \geq \lambda_p > 0 \quad (19)$$

for the eigenvalues, and an independent Haar measure on the Stiefel manifold, as the prior for the eigenvectors. The aforementioned prior on the eigenvalues has useful interesting properties which help to explain the type of eigenvalue truncation described in previous sections. We note that the prior is improper but the posterior is always proper.

Proposition 2

The prior on the eigenvalues in (19) is improper, whereas the posterior yields a proper distribution. More formally,

$$\int_C \pi(\underline{\lambda})d\underline{\lambda}=\int_C \exp(-g_{\max}\,\lambda_1/\lambda_p)d\underline{\lambda}=\infty,$$

and

$$\int_C \pi(\underline{\lambda})f(\underline{\lambda},\underline{l})d\underline{\lambda} \propto \int_C \exp(-(n/2)\sum_{i=1}^{p} l_i/\lambda_i)\,(\prod_{i=1}^{p}\lambda_i)^{-n/2}\exp(-g_{\max}\,\lambda_1/\lambda_p)d\underline{\lambda}<\infty,$$

where $\underline{\lambda} = (\lambda_1,\ldots,\lambda_p)$ *and* $C = \{\underline{\lambda} : \lambda_1 \geq \cdots \geq \lambda_p > 0\}$.

Proof—The proof is given in Supplemental Section A.

The prior above puts the greatest mass around the region $\{\underline{\lambda} \in \mathbb{R}^p : \lambda_1 = \cdots = \lambda_p\}$ which consequently encourages "shrinking" or "pulling" the eigenvalues closer together. Note that the support of both the prior and the posterior is the entire space of ordered eigenvalues. Hence the prior simply by itself does not immediately yield a hard constraint on the condition number. Evaluating the posterior mode yields an estimator that satisfies the condition number constraint.

A clear picture of the regularization achieved by the prior above and its potential for "eigenvalue shrinkage" emerges when compared to the other types of priors suggested in the literature and the corresponding Bayes estimators. The standard MLE implies of course a completely flat prior on the constrained space C. A commonly used inverse Wishart conjugate prior $\Sigma^{-1} \sim \text{Wishart}(m, cI)$ yields a posterior mode which is a linear shrinkage estimator (5) with $\delta = m/(n + m)$. Note however that the coefficients of the combination do not depend of the data X, and are a function only of the sample size n and the degrees of freedom or shape parameter from the prior, m. A useful prior for covariance matrices that yields a data-dependent posterior mode is the reference prior proposed by Yang and Berger (1994). For this prior, the eigenvalues are inversely proportional to the determinant of the the covariance matrix, as given by $\prod_{i=1}^{p} \lambda_i$, and also encourages shrinkage of the eigenvalues. The posterior mode using this reference prior can be formulated similarly to that of condition number regularization:

$$\arg\max_{\lambda_1 \geq \ldots \geq \lambda_p > 0} \exp(-(n/2)\sum_{i=1}^{p} l_i/\lambda_i)\,(\prod_{i=1}^{p}\lambda_i)^{-n/2}\,(\prod_{i=1}^{p}\lambda_i)^{-1} = \arg\min_{\lambda_1 \geq \ldots \geq \lambda_p > 0} (n/2)\sum_{i=1}^{p} l_i/\lambda_i + ((n+2)/2)\sum_{i=1}^{p}\log\lambda_i.$$

An examination of the penalty implied by the reference prior suggests that there is no direct penalty on the condition number. In Supplemental Section B we illustrate the density of the priors discussed above in the two-dimensional case. In particular, the density of the condition number regularization prior places more emphasis on the line $\lambda_1 = \lambda_2$ thus "squeezing" the eigenvalues together. This is in direct contrast with the inverse Wishart or reference priors where this shrinkage effect is not as pronounced.

5 Decision-theoretic risk properties

5.1 Asymptotic properties

We now show that the condition number-regularized estimator $\hat{\Sigma}_{\text{cond}}$ has asymptotically lower risk than the sample covariance matrix S with respect to entropy loss. Recall that the entropy loss, also known as Stein's loss, is defined as follows:

$$\mathscr{L}_{\text{ent}}(\widehat{\sum}, \sum) = \mathbf{tr}(\sum^{-1}\widehat{\sum}) - \log\det(\sum^{-1}\widehat{\sum}) - p. \quad (20)$$

Let $\lambda_1, \ldots, \lambda_p$, with $\lambda_1 \geq \cdots \geq \lambda_p$, denote the eigenvalues of the true covariance matrix Σ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$. Define $\underline{\lambda} = (\lambda_1, \ldots, \lambda_p)$, $\underline{\lambda}^{-1} = (\lambda_1^{-1}, \ldots, \lambda_p^{-1})$, and $\kappa = \lambda_1/\lambda_p$.

First consider the trivial case when $p > n$. In this case, the sample covariance matrix S is singular regardless of the singularity of Σ, and $\mathscr{L}_{\text{ent}}(S, \Sigma) = \infty$, whereas the loss and therefore the risk of $\hat{\Sigma}_{\text{cond}}$ are always finite. Thus, $\hat{\Sigma}_{\text{cond}}$ has smaller entropy risk than S.

For $p \leq n$, if the true covariance matrix has a finite condition number, it can be shown that for a properly chosen $\kappa_{\max}$, the condition number-regularized estimator asymptotically dominates the sample covariance matrix. This assertion is formalized below.

Theorem 3—Consider a class of covariance matrices $\mathcal{D}(\kappa, \omega)$, whose condition numbers are bounded above by κ and with minimum eigenvalue bounded below by $\omega > 0$, i.e.,

$$\mathcal{D}(\kappa, \omega) = \{\sum = R\Lambda R^T : R \text{ orthogonal}, \ \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p), \omega \leq \lambda_p \leq \ldots \leq \lambda_1 \leq \kappa\omega\}.$$

Then, the following results hold.

i. Consider the quantity $\tilde{\Sigma}(\kappa_{\max}, \omega) = Q\text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p)Q^T$, where

$$\tilde{\lambda}_i = \begin{cases} \omega, & \text{if } l_i \leq \omega \\ l_i, & \text{if } \omega \leq l_i < \kappa_{\max}\omega \\ \kappa_{\max}\omega, & \text{if } l_i \geq \kappa_{\max}\omega, \end{cases}$$

and the sample covariance matrix given as $S = Q\text{diag}(l_1, \ldots, l_p)Q^T$, Q orthogonal, $l_1 \geq \ldots \geq l_p$. If the true covariance matrix $\Sigma \in \mathcal{D}(\kappa_{\max}, \omega)$, then $\forall\, n$, $\hat{\Sigma}(\kappa_{\max}, \omega)$ has a smaller entropy risk than S.

ii. Consider a true covariance matrix Σ whose condition number is bounded above by κ, i.e., $\Sigma \in \cup_{\omega>0} \mathcal{D}(\kappa, \omega)$. If $\kappa_{\max} \geq \kappa(1 - \sqrt{\gamma})^{-2}$, then as $p/n \to \gamma \in (0, 1)$,

$$\mathbf{pr}\left(\left\{\sum \in \mathcal{D}(\kappa_{\max}, \tau^\star)\right\} \quad \text{eventually}\right) = 1,$$

where $\tau^\star = \tau^\star(\kappa_{\max})$ is given in (12).

Proof: The proof is given in Supplemental Section A.

Combining the two results above, we conclude that the estimator $\hat{\Sigma}_{\text{cond}} = \hat{\Sigma}(\kappa_{\max}, \tau^\star(\kappa_{\max}))$ asymptotically has a lower entropy risk than the sample covariance matrix.

5.2 Finite sample performance

This section undertakes a simulation study in order to compare the finite-sample risks of the condition number-regularized estimator $\hat{\Sigma}_{\text{cond}}$ with those of the sample covariance matrix (S) and the linear shrinkage estimator ($\hat{\Sigma}_{\text{LS}}$) in the "large p, small n" setting. The regularization parameter δ for $\hat{\Sigma}_{\text{LS}}$ is chosen as prescribed in Warton (2008). The optimal $\hat{\Sigma}_{\text{cond}}$ is calculated using the adaptive parameter selection method outlined in Section 3. Since $\hat{\Sigma}_{\text{cond}}$ and $\hat{\Sigma}_{\text{LS}}$ both select the regularization parameters similarly, *i.e.*, by minimizing the empirical predictive risk (18), a meaningful comparison between two estimators can be made. We consider two loss functions traditionally used in covariance estimation risk comparisons: (a) entropy loss as given in (20) and (b) quadratic loss

$$\mathcal{L}_{\text{Q}}(\widehat{\sum}, \sum) = \|\widehat{\sum}\sum^{-1} - I\|_{\text{F}}^2.$$

Condition number regularization applies shrinkage to both ends of the sample eigenvalue spectrum and does not affect the middle part, whereas linear shrinkage does this to the entire spectrum uniformly. Therefore, it is expected that $\hat{\Sigma}_{\text{cond}}$ works well when a small proportion

of eigenvalues are found at the extremes. Such situations rise very naturally when only a few eigenvalues explain most of the variation in data. To understand the performance of the estimators in this context the following scenarios were investigated. We consider diagonal matrices of dimensions $p = 120, 250, 500$ as true covariance matrices. The eigenvalues (diagonal elements) are dichotomous, where the "high" values are $(1 - \rho) + \rho p$ and the "low" values are $1 - \rho$. For each p, we vary the composition of the diagonal elements such that the high values take only one (singleton), 10%, 20%, 30%, and 40% of the total number of p eigenvalues. The sample size n is chosen so that $\gamma = p/n$ is approximately 1.25, 2, or 4. Note that for a given p, the condition number of the true covariance matrices is held fixed at $1 + p\rho/(1 - \rho)$ regardless of the composition of eigenvalues. For each of the simulation scenarios, we generate 1000 data sets and compute 1000 estimates of the true covariance matrix. The risks are calculated by averaging the losses over these 1000 estimates.

Figure 4 presents the results for $\rho = 0.1$, and represents a large condition number. It is observed that in general $\hat{\Sigma}_{\text{cond}}$ has less risk than $\hat{\Sigma}_{\text{LS}}$, which in turn has less risk than the sample covariance matrix (entropy loss is not defined for the sample covariance matrix). This phenomenon is most clear when the eigenvalue spectrum has a singleton in the extreme. In this case, $\hat{\Sigma}_{\text{cond}}$ gives a risk reduction between 27 % and 67 % for entropy loss, and between 67 % and 91 % for quadratic loss. The risk reduction tends to be more pronounced in high dimensional scenarios, *i.e.*, for p large and n small. The performance of $\hat{\Sigma}_{\text{cond}}$ over $\hat{\Sigma}_{\text{LS}}$ is maintained until the "high" eigenvalues compose up to 30 % of the eigenvalue spectrum. Comparing the two loss functions, risk reduction of $\hat{\Sigma}_{\text{cond}}$ is more distinct in quadratic loss. We note that for quadratic loss with large p and large proportion of "high" eigenvalues, there are cases that the sample covariance matrix can perform well.

As an example of a moderate condition number, results for the $\rho = 0.5$ case is given in Supplemental Section C. General trends are the same as with the $\rho = 0.1$ case.

In summary, the risk comparison study provides a numerical evidence that condition number regularization has merit when the true covariance matrix has a bimodal eigenvalue distribution and/or the true condition number is large.

6 Application to portfolio selection

This section illustrates the merits of the condition number regularization in the context of financial portfolio optimization, where a well-conditioned covariance estimator is necessary. A portfolio refers to a collection of risky assets held by an institution or an individual. Over the holding period, the return on the portfolio is the weighted average of the returns on the individual assets that constitutes the portfolio, in which the weight associated with each asset corresponds to its proportion in monetary terms. The objective of portfolio optimization is to determine the weights that maximize the return on the portfolio. Since the asset returns are stochastic, a portfolio always carries a risk of loss. Hence the objective is to maximize the overall return subject to a given level of risk, or equivalently to minimize risk for a given level of return. Mean-variance portfolio (MVP) theory (Markowitz, 1952) uses the standard deviation of portfolio returns to quantify the risk. Estimation of the covariance matrix of asset returns thus becomes critical in the MVP setting. An important and difficult component of MVP theory is to estimate the expected return on the portfolio (Luenberger, 1998; Merton, 1980). Since the focus of this paper lies in estimating covariance matrices and not expected returns, we shall focus on determining the *minimum* variance portfolio which only requires an estimate of the covariance matrix; see, *e.g.*, Chan et al. (1999). For this we shall use the condition number regularization, linear shrinkage and the sample covariance matrix, in constructing a minimum variance portfolio. We compare their respective performance over a period of more than 14 years.

6.1 Minimum variance portfolio rebalancing

We begin with a formal description of the minimum variance portfolio selection problem. The universe of assets consists of p risky assets, denoted $1,\ldots, p$. We use r_i to denote the return of asset i over one period; that is, its change in price over one time period divided by its price at the beginning of the period. Let Σ denote the covariance matrix of $r = (r_1,\ldots, r_p)$. We employ w_i to denote the weight of asset i in the portfolio held throughout the period. A long position in asset i corresponds to $w_i > 0$, and a short position corresponds to $w_i < 0$. The portfolio is therefore unambiguously represented by the vector of weights $w = (w_1,\ldots, w_p)$. Without loss of generality, the budget constraint can be written as $\mathbf{1}^T w = 1$, where $\mathbf{1}$ is the vector of all ones. The risk of a portfolio is measured by the standard deviation $(w^T \Sigma w)^{1/2}$ of its return.

Now the minimum variance portfolio selection problem can be formulated as

$$\begin{array}{ll} \text{minimize} & w^T \sum w \\ \text{subject to} & \mathbf{1}^T w = 1. \end{array} \quad (21)$$

This is a simple quadratic program that has an analytic solution $w^\star = (\mathbf{1}^T \Sigma^{-1} \mathbf{1})^{-1} \Sigma^{-1} \mathbf{1}$. In practice, the parameter Σ has to be estimated.

The standard portfolio selection problem described above assumes that the returns are stationary, which is of course not realistic. As a way of dealing with the nonstationarity of returns, we employ a minimum variance portfolio rebalancing (MVR) strategy as follows.

Let $r^{(t)} = (r_1^{(t)}, \ldots, r_p^{(t)}) \in \mathbb{R}^p$, $t = 1, \ldots, N_{\text{tot}}$, denote the realized returns of assets at time t (the time unit under consideration can be a day, a week, or a month). The periodic minimum variance rebalancing strategy is implemented by updating the portfolio weights every L time units, i.e., the entire trading horizon is subdivided into blocks each consisting of L time units. At the start of each block, we determine the minimum variance portfolio weights based on the past N_{estim} observations of returns. We shall refer to N_{estim} as the estimation horizon size. The portfolio weights are then held constant for L time units during these "holding" periods, i.e., during each of these blocks, and subsequently updated at the beginning of the following one. For simplicity, we shall assume the entire trading horizon consists of $N_{\text{tot}} = N_{\text{estim}} + KL$ time units, for some positive integer K, *i.e.*, there will be K updates. (The last rebalancing is done at the end of the entire period, and so the out-of-sample performance of the rebalanced portfolio for this holding period is not taken into account.) We therefore have a series of portfolios $w^{(j)} = (\mathbf{1}^T (\hat{\Sigma}^{(j)})^{-1} \mathbf{1})^{-1} (\hat{\Sigma}^{(j)})^{-1} \mathbf{1}$ over the holding periods of $[N_{\text{estim}} + 1 + (j-1)L, N_{\text{estim}} + jL]$, $j = 1,\ldots, K$. Here $\hat{\Sigma}^{(j)}$ is the covariance matrix of the asset returns estimated from those over the jth holding period.

6.2 Empirical out-of-sample performance

In this empirical study, we use the 30 stocks that constituted the Dow Jones Industrial Average as of July 2008 (Supplemental Section D.1 lists these 30 stocks). We used the closing prices adjusted daily for all applicable splits and dividend distributions downloaded from Yahoo! Finance (http://finance.yahoo.com/). The whole period considered in our numerical study is from the trading date of December 14, 1992 to June 6, 2008 (this period consists of 4100 trading days). We consider weekly returns: the time unit is 5 consecutive trading days. We take

$$N_{\text{tot}} = 820, \quad L = 15, \quad N_{\text{estim}} = 15, 30, 45, 60.$$

To estimate the covariance matrices, we use the last N_{estim} weekly returns of the constituents of the Dow Jones Industrial Average.[3] The entire trading horizon corresponds to $K = 48$ holding periods, which span the dates from February 18, 1994 to June 6, 2008. In what follows, we compare the MVR strategy where the covariance matrices are estimated using the condition number regularization with those that use either the sample covariance matrix or linear shrinkage. We employ two linear shrinkage schemes: that of Warton (2008) of Section 5.2 and that of Ledoit and Wolf (2004). The latter is widely accepted as a well-conditioned estimator in the financial literature.

Performance metrics—We use the following quantities in assessing the performance of the MVR strategies. For precise formulae of these metrics, refer to Supplemental Section D.3.

- *Realized return.* The realized return of the portfolio over the trading period.
- *Realized risk.* The realized risk (return standard deviation) of the portfolio over the trading period.
- *Realized Sharpe ratio.* The realized excess return, with respect to the risk-free rate, per unit risk of the portfolio.
- *Turnover.* Amount of new portfolio assets purchased or sold over the trading period.
- *Normalized wealth growth.* Accumulated wealth yielded by the portfolio over the trading period when the initial budget is normalized to one, taking the transaction cost into account.
- *Size of the short side.* The proportion of the short side (negative) weights to the sum of the absolute weights of the portfolio.

We assume that the transaction costs are the same for the 30 stocks and set them to 30 basis points. The risk-free rate is set at 5% per annum.

Comparison results—Figure 5 shows the normalized wealth growth over the trading horizon for four different values of N_{estim}. The sample covariance matrix failed in solving (21) for $N_{\text{estim}} = 15$ because of its singularity and hence is omitted in this figure. The MVR strategy using the condition number-regularized covariance matrix delivers higher growth as compared to using the sample covariance matrix, linear shrinkage or index tracking in this performance metric. The higher growth is realized consistently across the 14 year trading period and is regardless of the estimation horizon. A trading strategy based on the condition number-regularized covariance matrix consistently performs better than the S&P 500 index and can lead to profits as much as 175% more than its closest competitor.

A useful result appears after further analysis. There is no significant difference between the condition number regularization approach and the two linear shrinkage schemes in terms of the realized return, risk, and Sharpe ratio. Supplemental Section D.4 summarizes these metrics for each estimator respectively averaged over the trading period. For all values of N_{estim}, the average differences of the metrics between the two regularization schemes are within two standard errors of those. Hence the condition number regularized estimator delivers better normalized wealth growth than the other estimators but without compromising on other measures such as volatility.

[3]Supplemental Section D.2 shows the periods determined by the choice of the parameters.

The turnover of the portfolio seems to be one of the major driving factors of the difference in wealth growth. In particular, the MVR strategy using the condition number-regularized covariance matrix gives far lower turnover and thus more stable weights than when using the linear shrinkage estimator or the sample covariance matrix. (See Supplemental Section D.5 for plots.) A lower turnover also implies less transaction costs, thereby also partially contributing to the higher wealth growth. Note that there is no explicit limit on turnover. The stability of the MVR portfolio using the condition number regularization appears to be related to its small size of the short side (reported in Supplemental Section D.6). Because stock borrowing is expensive, the condition number regularization based strategy can be advantageous in practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the editor and the associate editor for useful comments that improved the presentation of the paper. J. Won was partially supported by the US National Institutes of Health (NIH) (MERIT Award R37EB02784) and by the US National Science Foundation (NSF) grant CCR 0309701. J. Lim's research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number: 2010-0011448). B. Rajaratnam was supported in part by NSF under grant nos. DMS-09-06392, DMS-CMG 1025465, AGS-1003823, DMS-1106642 and grants NSA H98230-11-1-0194, DARPA-YFA N66001-11-1-4131, and SUWIEVP10-SUFSC10-SMSCVISG0906.

References

Banerjee O, El Ghaoui L, D'Aspremont A. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. Journal of Machine Learning Research. 2008; 9:485–516.

Boyd, S.; Vandenberghe, L. Convex Optimization. Cambridge University Press; 2004.

Chan N, Karceski N, Lakonishok J. On portfolio optimization: Forecasting covariances and choosing the risk model. Review of Financial Studies. 1999; 12(5):937–974.

Daniels M, Kass R. Shrinkage estimators for covariance matrices. Biometrics. 2001; 57:1173–1184. [PubMed: 11764258]

Dempster AP. Covariance Selection. Biometrics. 1972; 28(1):157–175.

Dey DK, Srinivasan C. Estimation of a covariance matrix under Stein's loss. The Annals of Statistics. 1985; 13(4):1581–1591.

Farrell, RH. Multivariate calculation. Springer-Verlag; New York: 1985.

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. [PubMed: 18079126]

Haff LR. The variational form of certain Bayes estimators. The Annals of Statistics. 1991; 19(3):1163–1190.

Hero A, Rajaratnam B. Large-scale correlation screening. Journal of the American Statistical Association. 2011; 106(496):1540–1552.

Hero A, Rajaratnam B. Hub discovery in partial correlation graphs. Information Theory, IEEE Transactions on. 2012; 58(9):6064–6078.

James, W.; Stein, C. Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability; Stanford, California, United States. 1961. p. 361-379.

Khare K, Rajaratnam B. Wishart distributions for decomposable covariance graph models. The Annals of Statistics. 2011; 39(1):514–555.

Ledoit O, Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance. 2003 Dec; 10(5):603–621.

Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis. 2004; 88:365–411.

Ledoit O, Wolf M. Nonlinear shrinkage estimation of large-dimensional covariance matrices. The Annals of Statistics. 2012 Jul; 40(2):1024–1060.

Letac G, Massam H. Wishart distributions for decomposable graphs. The Annals of Statistics. 2007; 35(3):1278–1323.

Lin S, Perlman M. A Monte-Carlo comparison of four estimators of a covariance matrix. Multivariate Analysis. 1985; 6:411–429.

Luenberger, DG. Investment science. Oxford University Press; New York: 1998.

Markowitz H. Portfolio selection. Journal of Finance. 1952; 7(1):77–91.

Merton R. On estimating expected returns on the market: An exploratory investigation. Journal of Financial Economics. 1980; 8:323–361.

Michaud RO. The Markowitz Optimization Enigma: Is Optimized Optimal. Financial Analysts Journal. 1989; 45(1):31–42.

Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association. 2009; 104(486):735–746. [PubMed: 19881892]

Pourahmadi M, Daniels MJ, Park T. Simultaneous modelling of the cholesky decomposition of several covariance matrices. Journal of Multivariate Analysis. 2007 Mar; 98(3):568–587.

Rajaratnam B, Massam H, Carvalho C. Flexible covariance estimation in graphical Gaussian models. The Annals of Statistics. 2008; 36(6):2818–2849.

Sheena Y, Gupta A. Estimation of the multivariate normal covariance matrix under some restrictions. Statistics & Decisions. 2003; 21:327–342.

Stein, C. Technical Report 6. Dept. of Statistics, Stanford University; 1956. Some problems in multivariate analysis Part I.

Stein, C. Estimation of a covariance matrix. Reitz Lecture, IMS-ASA Annual Meeting; 1975.

Stein C. Lectures on the theory of estimation of many parameters (English translation). Journal of Mathematical Sciences. 1986; 34(1):1373–1403.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996; 58(1):267–288.

Warton DI. Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices. Journal of the American Statistical Association. 2008; 103(481):340–349.

Won, JH.; Kim, S-J. Maximum Likelihood Covariance Estimation with a Condition Number Constraint. Proceedings of the Fortieth Asilomar Conference on Signals, Systems and Computers; 2006. p. 1445-1449.

Yang R, Berger JO. Estimation of a covariance matrix using the reference prior. The Annals of Statistics. 1994; 22(3):1195–1211.

Appendix

Proof of Lemma 1

Recall the spectral decomposition of the sample covariance matrix $S = QLQ^T$, with $L = \text{diag}(l_1,\ldots, l_p)$ and $l_1 \geq \ldots \geq l_p \geq 0$. From the objective function in (8), suppose the variable Ω has the spectral decomposition RMR^T, with R orthogonal and $M = \text{diag}(\mu_1,\ldots, \mu_p)$, $\mu_1 \leq \ldots \leq \mu_p$. Then the objective function in (8) can be written as

$$\begin{aligned}\mathbf{tr}(\Omega S) - \log\det(\Omega) &= \mathbf{tr}(RMR^T QLQ^T) - \log\det(RMR^T) \\ &= l(R, M) \\ &\geq \mathbf{tr}(ML) - \log\det M = l(Q, M),\end{aligned}$$

with equality in the last line when $R = Q$ (Farrell, 1985, Ch. 14). Hence (8) amounts to

$$\begin{array}{ll}\text{minimize} & \sum_{i=1}^{p}(l_i\mu_i - \log\mu_i) \\ \text{subject to} & u \le \mu_1 \le \cdots \le \mu_p \le \kappa_{\max} u, \quad i=1,\ldots,p,\end{array} \tag{22}$$

with the optimization variables $\mu_1, \ldots, \mu_p$ and u.

For the moment we shall ignore the order constraints among the eigenvalues. Then problem (22) becomes separable in $\mu_1,\ldots,\mu_p$. Call this related problem (22*). For a fixed u, the minimizer of each summand of the objective function in (22) without the order constraints is given as

$$\mu_i^\star(u) = \underset{u \le \mu_i \le \kappa_{\max} u}{\operatorname{argmin}} \; (l_i\mu_i - \log\mu_i) = \min\{\max\{u, 1/l_i\}, \kappa_{\max} u\}. \tag{23}$$

Note that (23) however satisfies the order constraints. In other words, $\mu_1^\star(u) \le \ldots \le \mu_p^\star(u)$ for all u. Therefore (22*) is equivalent to (22). Plugging (23) in (22) removes the constraints and the objective function reduces to a univariate one:

$$J_{\kappa_{\max}}(u) \triangleq \sum_{i=1}^{p} J_{\kappa_{\max}}^{(i)}(u), \tag{24}$$

where

$$J_{\kappa_{\max}}^{(i)}(u) = l_i\mu_i^\star(u) - \log\mu_i^\star(u) = \begin{cases} l_i(\kappa_{\max}u) - \log(\kappa_{\max}u), & u < 1/(\kappa_{\max}l_i) \\ 1 + \log l_i, & 1/(\kappa_{\max}l_i) \le u \le 1/l_i \\ l_i u - \log u, & u > 1/l_i. \end{cases}$$

The function (24) is convex, since each $J_{\kappa_{\max}}^{(i)}$ is convex in u.

Proof of Theorem 1

The function $J_{\kappa_{\max}}^{(i)}(u)$ is convex and is constant on the interval $[1/(\kappa_{\max}l_i), 1/l_i]$. Thus, the function $J_{\kappa_{\max}}(u) \triangleq \sum_{i=1}^{p} J_{\kappa_{\max}}^{(i)}(u)$ has a region on which it is a constant if and only if

$$[1/(\kappa_{\max}l_1), 1/l_1] \cap \left[1/(\kappa_{\max}l_p), 1/l_p\right] \neq \emptyset,$$

or equivalently, $\kappa_{\max} > \operatorname{cond}(S)$. Therefore, provided that $\kappa_{\max} \le \operatorname{cond}(S)$, the convex function $J_{\kappa\max}(u)$ does not have a region on which it is constant. Since $J_{\kappa\max}(u)$ is strictly decreasing for $0 < u < 1/(\kappa_{\max}l_1)$ and strictly increasing for $u > 1/l_p$, it has a unique minimizer $u^\star$. If $\kappa_{\max} > \operatorname{cond}(S)$, the maximizer $u^\star$ may not be unique because $J_{\kappa\max}(u)$ has a plateau. However, since the condition number constraint becomes inactive, $\hat{\Sigma}_{\text{cond}} = S$ for all the maximizers.

Now assume that $\kappa_{\max} \le \operatorname{cond}(S)$. For $\alpha \in \{1,\ldots, p-1\}$ and $\beta \in \{2, \ldots, p\}$, define the following two quantities.

$$u_{\alpha,\beta}=\frac{\alpha+p-\beta+1}{\sum_{i=1}^{\alpha} l_i+\sum_{i=\beta}^{p}\kappa_{\max} l_i} \text{ and } v_{\alpha,\beta}=\kappa_{\max} u_{\alpha,\beta}.$$

By construction, $u_{\alpha,\beta}$ coincides with $u^\star$ if and only if

$$1/l_\alpha<u_{\alpha,\beta} \le 1/l_{\alpha+1} \text{ and } 1/l_{\beta-1} \le \kappa_{\max} u_{\alpha,\beta}<1/l_\beta. \quad (25)$$

Consider a set of rectangles $\{R_{\alpha,\beta}\}$ in the uv-plane such that

$$R_{\alpha,\beta}=\{(u,v):1/l_\alpha<u \le 1/l_{\alpha+1} \text{ and } 1/l_{\beta-1} \le v<1/l_\beta\}.$$

Then condition (25) is equivalent to

$$(u_{\alpha,\beta}, v_{\alpha,\beta}) \in R_{\alpha,\beta} \quad (26)$$

in the uv-plane. Since $\{R_{\alpha,\beta}\}$ partitions $\{(u, v) : 1/l_1 < u \le 1/l_p \text{ and } 1/l_1 \le v < 1/l_p\}$ and $(u_{\alpha,\beta}, v_{\alpha,\beta})$ is on the line $v = \kappa_{\max} u$, an obvious algorithm to find the pair (α, β) that satisfies the condition (26) is to keep track of the rectangles $R_{\alpha,\beta}$ that intersect this line. To understand that algorithm takes $O(p)$ operations, start from the origin of the uv-plane, increase u and v along the line $v = \kappa_{\max} u$. Since $\kappa_{\max} > 1$, if the line intersects $R_{\alpha,\beta}$, then the next intersection occurs in one of the three rectangles: $R_{\alpha+1,\beta}$, $R_{\alpha,\beta+1}$, and $R_{\alpha+1,\beta+1}$. Therefore after finding the first intersection (which is on the line $u = 1/l_1$), the search requires at most $2p$ tests to satisfy condition (26). Finding the first intersection takes at most p tests.

Proof of Proposition 1

Recall that, for $\kappa_{\max} = \nu_0$,

$$u^\star(\nu_0)=\frac{\alpha+p-\beta+1}{\sum_{i=1}^{\alpha} l_i+\nu_0\sum_{i=\beta}^{p} l_i}$$

and

$$v^\star(\nu_0)=\nu_0 u^\star(\nu_0)=\frac{\alpha+p-\beta+1}{\frac{1}{\nu_0}\sum_{i=1}^{\alpha} l_i+\sum_{i=\beta}^{p} l_i},$$

where $\alpha = \alpha(\nu_0) \in \{1,\ldots, p\}$ is the largest index such that $1/l_\alpha < u^\star(\nu_0)$ and $\beta = \beta(\nu_0) \in \{1, \ldots, p\}$ is the smallest index such that $1/l_\beta > \nu_0 u^\star(\nu_0)$. Then

$$1/l_\alpha<u^\star(\nu_0) \le 1/l_{\alpha+1}$$

and

$$1/l_{\beta-1} \le v^\star(\nu_0)<1/l_\beta.$$

The lower and upper bounds $u^\star(\nu_0)$ and $v^\star(\nu_0)$ of the reciprocal sample eigenvalues can be divided into four cases:

1. $1/l_\alpha < u^\star(\nu_0) < 1/l_{\alpha+1}$ and $1/l_{\beta-1} < v^\star(\nu_0) < 1/l_\beta$.

We can find $\nu > \nu_0$ such that

$$1/l_\alpha<u^\star(\nu) \le 1/l_{\alpha+1}$$

and

$$1/l_{\beta-1} \le v^\star(\nu)<1/l_\beta.$$

Therefore,

$$u^\star(\nu)=\frac{\alpha+p-\beta+1}{\sum_{i=1}^{\alpha} l_i+\nu\sum_{i=\beta}^{p} l_i}<\frac{\alpha+p-\beta+1}{\sum_{i=1}^{\alpha} l_i+\nu_0\sum_{i=\beta}^{p} l_i}=u^\star(\nu_0)$$

and

$$v^\star(\nu)=\frac{\alpha+p-\beta+1}{\frac{1}{\nu_0}\sum_{i=1}^{\alpha} l_i+\sum_{i=\beta}^{p} l_i}>\frac{\alpha+p-\beta+1}{\frac{1}{\nu}\sum_{i=1}^{\alpha} l_i+\sum_{i=\beta}^{p} l_i}=v^\star(\nu_0).$$

2. $u^\star(\nu_0) = 1/l_{\alpha+1}$ and $1/l_{\beta-1} < v^\star(\nu_0) < 1/l_\beta$.

Suppose $u^\star(\nu) > u^\star(\nu_0)$. Then we can find $\nu > \nu_0$ such that $\alpha(\nu) = \alpha(\nu_0) + 1 = \alpha + 1$ and $\beta(\nu) = \beta(\nu_0) = \beta$. Then,

$$u^\star(\nu)=\frac{\alpha+1+p-\beta+1}{\sum_{i=1}^{\alpha+1} l_i+\nu\sum_{i=\beta}^{p} l_i}.$$

Therefore,

$$\begin{aligned}\frac{1}{u^\star(\nu_0)}-\frac{1}{u^\star(\nu)}&=1/l_{\alpha+1}-\frac{\sum_{i=1}^{\alpha+1} l_i+\nu\sum_{i=\beta}^{p} l_i}{\alpha+1+p-\beta+1}\\&=\frac{(\alpha+p-\beta+1)l_{\alpha+1}-(\sum_{i=1}^{\alpha+1} l_i+\nu\sum_{i=\beta}^{p} l_i)}{\alpha+1+p-\beta+1}>0,\end{aligned}$$

or

$$l_{\alpha+1}>\frac{\sum_{i=1}^{\alpha+1} l_i+\nu\sum_{i=\beta}^{p} l_i}{\alpha+p-\beta+1}>\frac{\sum_{i=1}^{\alpha+1} l_i+\nu_0\sum_{i=\beta}^{p} l_i}{\alpha+p-\beta+1}=l_{\alpha+1},$$

which is a contradiction. Therefore, $u^\star(\nu) \le u^\star(\nu_0)$.

Now, we can find $\nu > \nu_0$ such that $\alpha(\nu) = \alpha(\nu_0) = \alpha$ and $\beta(\nu) = \beta(\nu_0) = \beta$. This reduces to case 1.

3. $1/l_\alpha < u^\star(\nu_0) < 1/l_{\alpha+1}$ and $v^\star(\nu_0) = 1/l_{\beta-1}$.

Suppose $v^\star(\nu) < v^\star(\nu_0)$. Then we can find $\nu > \nu_0$ such that $\alpha(\nu) = \alpha(\nu_0) = \alpha$ and $\beta(\nu) = \beta(\nu_0) - 1 = \beta - 1$. Then,

$$v^\star(\nu) = \frac{\alpha + p - \beta + 2}{\frac{1}{\nu}\sum_{i=1}^{\alpha} l_i + \sum_{i=\beta-1}^{p} l_i}.$$

Therefore,

$$\frac{1}{v^\star(\nu_0)} - \frac{1}{v^\star(\nu)} = 1/l_{\beta-1} - \frac{\sum_{i=1}^{\alpha} l_i + \nu\sum_{i=\beta-1}^{p} l_i}{\alpha + p - \beta + 2} = \frac{(\alpha + p - \beta + 1)l_{\beta-1} - (\sum_{i=1}^{\alpha} l_i + \nu\sum_{i=\beta}^{p} l_i)}{\alpha + p - \beta + 2} < 0,$$

or

$$l_{\beta-1} < \frac{\sum_{i=1}^{\alpha} \frac{1}{\nu} l_i + \sum_{i=\beta-1}^{p} l_i}{\alpha + p - \beta + 1} < \frac{\sum_{i=1}^{\alpha+1} \frac{1}{\nu_0} l_i + \sum_{i=\beta}^{p} l_i}{\alpha + p - \beta + 1} = l_{\beta-1},$$

which is a contradiction. Therefore, $v^\star(\nu) \geq v^\star(\nu_0)$.

Now, we can find $\nu > \nu_0$ such that $\alpha(\nu) = \alpha(\nu_0) = \alpha$ and $\beta(\nu) = \beta(\nu_0) = \beta$. This reduces to case 1.

4. $u^\star(\nu_0) = 1/l_{\alpha+1}$ and $v^\star(\nu_0) = 1/l_{\beta-1}$. $1/l_{\alpha+1} = u^\star(\nu_0) = v^\star(\nu_0)/\nu_0 = 1/(\nu_0 l_{\beta-1})$. This is a measure zero event and does not affect the conclusion.

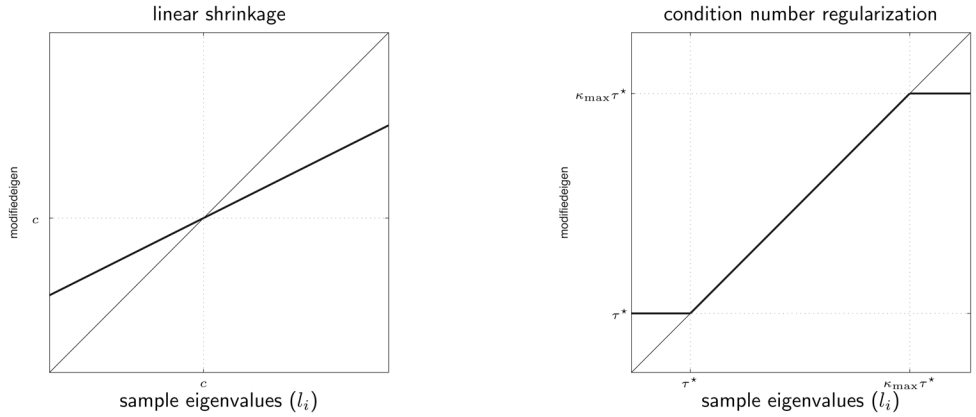

Figure 1.
Comparison of eigenvalue shrinkage of the linear shrinkage estimator (left) and the condition number-constrained estimator (right).

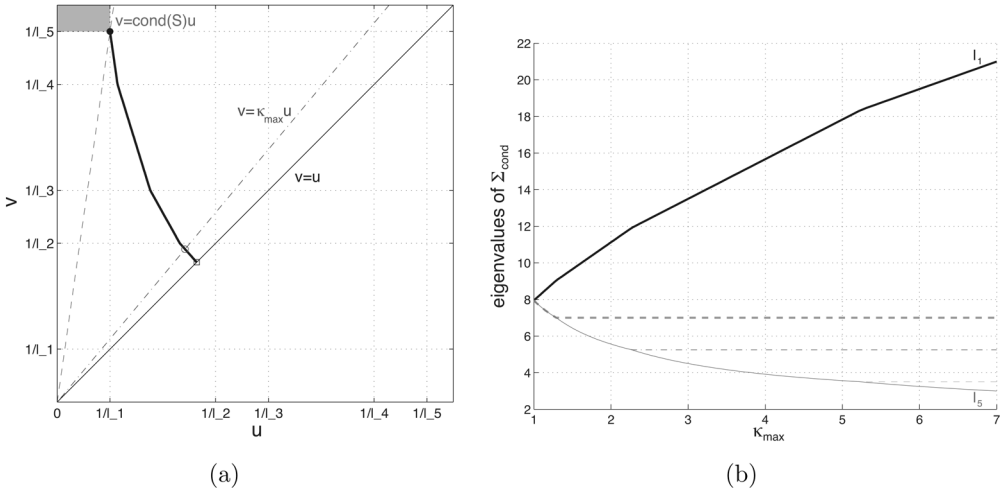

Figure 2.

Regularization path of the condition number constrained estimator. (a) Path of $(u^\star(\kappa_{\max}), v^\star(\kappa_{\max}))$ on the u-v plane, for sample eigenvalues (21, 7, 5.25, 3.5, 3) (thick curve). (b) Regularization path of the same sample eigenvalues as a function of $\kappa_{\max}$. Note that the estimates decreases monotonically as the condition number constraint $\kappa_{\max}$ decreases to 1.

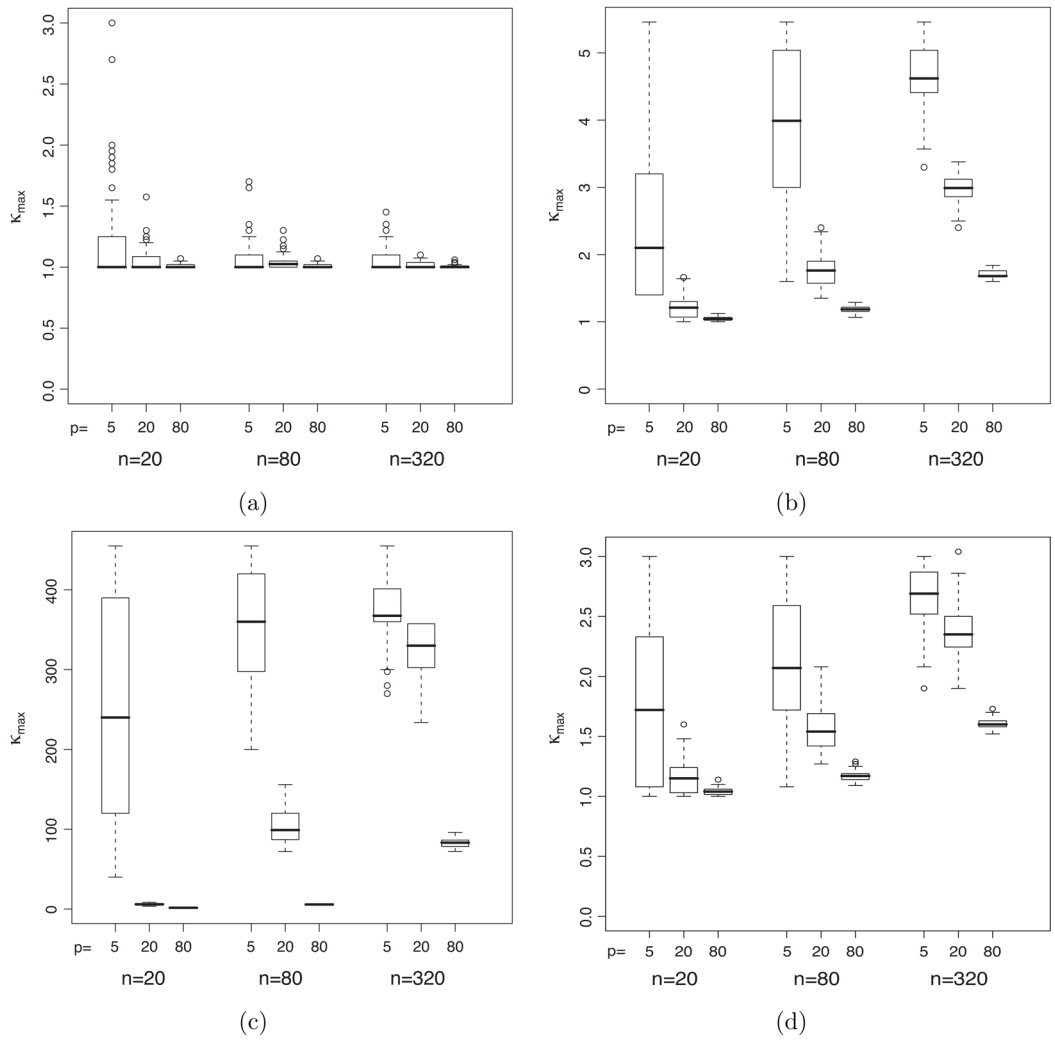

Figure 3.

Box plots summarizing the distribution of $\hat{\kappa}_{\max}$ for dimensions $p = 5, 20, 80$ and for sample sizes $n = 20, 80, 320$ for the following covariance matrices (a) identity (b) diagonal exponentially decreasing, condition number 5, (c) diagonal exponentially decreasing, condition number 400, (d) Toeplitz matrix whose (i, j)th element is $0.3^{|i-j|}$ for $i, j = 1, 2, \ldots, p$.

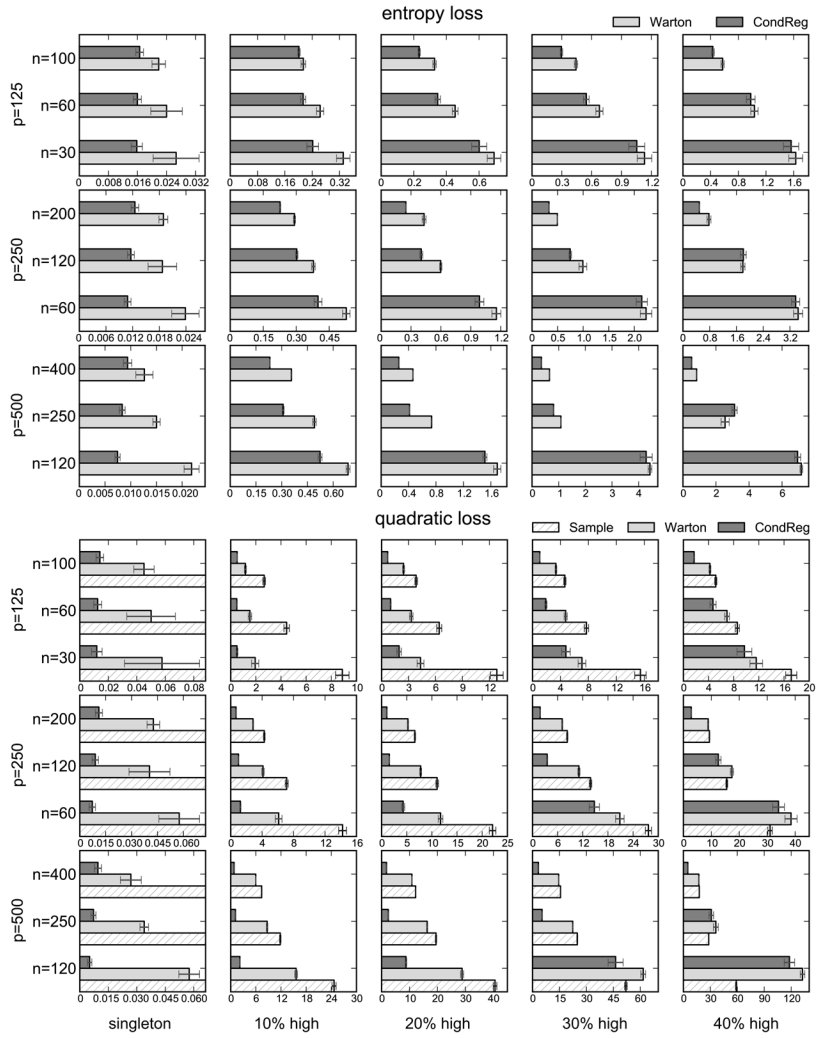

Figure 4.

Average risks (with error bars) over 1000 runs with respect to two loss functions when ρ= 0.1. sample=sample covariance matrix, Warton=linear shrinkage (Warton, 2008), CondReg=condition number regularization. Risks are normalized by the dimension (p).

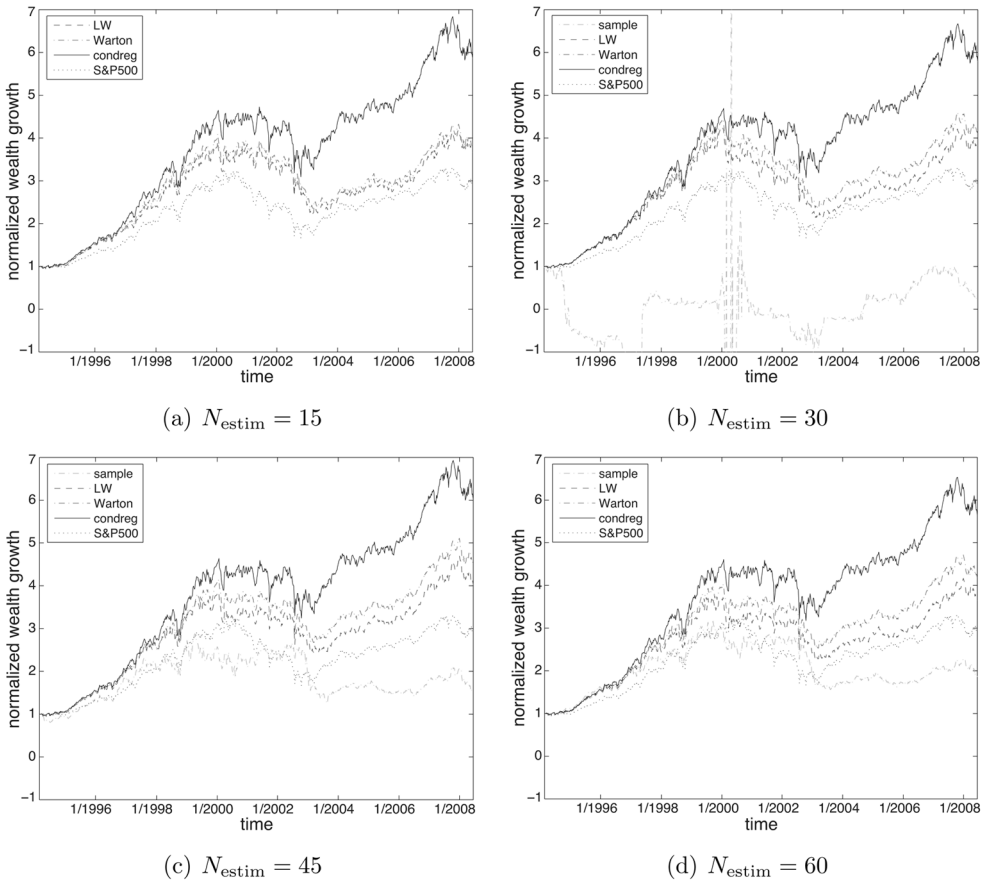

Figure 5.

Normalized wealth growth results of the minimum variance rebalancing strategy for various estimation horizon sizes over the trading period from February 18, 1994 through June 6, 2008. sample=sample covariance matrix, LW=linear shrinkage (Ledoit and Wolf, 2004), Warton=linear shrinkage (Warton, 2008), condreg=condition number regularization. For comparison, the S&P 500 index for the same period (i.e., index tracking), with the initial price normalized to 1, is also plotted.