MAXIMUM LIKELIHOOD COVARIANCE ESTIMATION WITH A CONDITION NUMBER CONSTRAINT

By

Joong-Ho Won
Johan Lim
Seung-Jean Kim
Bala Rajaratnam

Technical Report No. 2009-10
August 2009

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

MAXIMUM LIKELIHOOD COVARIANCE ESTIMATION WITH A CONDITION NUMBER CONSTRAINT

By

Joong-Ho Won
Department of Health Research and Policy
Stanford University

Johan Lim
Department of Statistics
Seoul National University

Seung-Jean Kim
Citi Alternative Investments
New York City

Bala Rajaratnam
Department of Statistics
Stanford University

Technical Report No. 2009-10
August 2009

This research was supported in part by National Science Foundation grant DMS 0505303.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Maximum Likelihood Covariance Estimation with a Condition Number Constraint

Joong-Ho Won* Johan Lim† Seung-Jean Kim‡ Bala Rajaratnam§

July 24, 2009

Abstract

High dimensional covariance estimation is known to be a difficult problem, has many applications and is of current interest to the larger statistical community. We consider the problem of estimating the covariance matrix of a multivariate normal distribution in the "large p small n" setting. Several approaches to high dimensional covariance estimation have been proposed in the literature. In many applications, the covariance matrix estimate is required to be not only invertible but also well-conditioned. Although many estimators attempt to do this, none of them address this problem directly. In this paper, we propose a maximum likelihood approach with an explicit constraint on the condition number to try and obtain a well-conditioned estimator. We demonstrate that the proposed estimation approach is computationally efficient, can be interpreted as a type of nonlinear shrinkage estimator, and has a natural Bayesian interpretation. We fully investigate the theoretical properties of the proposed estimator and proceed to develop an approach that adaptively determines the level of regularization that is required. Finally we investigate the performance of the estimator in simulated and real-life examples and demonstrate that it has good risk properties and can serve as a competitive procedure especially when the sample size is small and when a well-conditioned estimator is required.

1 Introduction

We consider the problem of estimation of the covariance matrix Σ of an p-dimensional multivariate Gaussian model. Since the seminal work of Stein (1975) and Dempster (1972) the

*Division of Biostatistics, Department of Health Research and Policy, Stanford University, Stanford, CA 94305, U.S.A. Email: jhwon@stanford.edu. Supported partially by NSF grant CCR 0309701 and NIH MERIT Award R37EB02784.

†Department of Statistics, Seoul National University, Seoul, Korea. Email: johanlim@snu.ac.kr.

‡Citi Alternative Investments, New York City, NY, U.S.A. Email: seungjean@gmail.com.

§Department of Statistics, Stanford University, CA 94305, U.S.A. Email: brajarat@stanford.edu. Supported in part by NSF grant DMS 0505303.

problem of estimating Σ is recognized as highly challenging. Formally, given n independent samples $x_1, \cdots, x_n \in \mathbb{R}^p$ from a zero-mean p-dimensional Gaussian distribution with an unknown covariance matrix Σ, the log-likelihood function of the covariance matrix has the form

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

where both $|\Sigma|$ and $\det\Sigma$ denote the determinant of Σ, $\mathbf{Tr}(A)$ denotes the trace of A, and S is the sample covariance matrix, *i.e.*,

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

The log-likelihood function is maximized by the sample covariance, *i.e.*, the maximum likelihood estimate (MLE) of the covariance is S (Anderson, 1970).

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 number of parameters. When $n < p$ the sample covariance matrix S is singular and not positive definite and hence it cannot be inverted to compute the precision matrix (the inverse of the covariance matrix), which is also needed in many applications. However, even when $n > p$, the eigenstructure tends to be systematically distorted unless p/n is extremely small, resulting in ill-conditioned estimators for Σ; see Dempster (1972) and Stein (1975).

Numerous papers have explored better alternative estimators for Σ (or Σ^{-1}) in both the frequentist and Bayesian frameworks. Many of these estimators give substantial risk reductions compared to the sample covariance estimator S in small sample sizes. A common underlying property of many of these estimators is that they are shrinkage estimators in the sense of James-Stein (James and Stein, 1961; Stein, 1956). A simple example is a family of linear shrinkage estimators which take a convex combination of the sample covariance and a suitably chosen target or regularization matrix. Ledoit and Wolf (2004b) study a linear shrinkage estimator towards a specified target covariance matrix, and choose the optimal shrinkage to minimize the Frobenius risk. Warton (2008) minimizes the predictive risk which is estimated using a cross-validation method, and studies its application to testing equality of means of two populations. Many other James-Stein type shrinkage estimators have been proposed and analyzed from a decision-theoretic point of view. To list a few, James and Stein (1961) study a constant risk minimax estimator and its modification in a class of orthogonally invariant estimators (we use one in our numerical study in Section 5.2). Dey and Srinivasan (1985) provide another minimax estimator which dominates the James-Stein estimator. Bayesian approaches often directly yield estimators which "shrink" towards 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 the eigenvalues of Σ; see, *e.g.*, Haff (1991). Yang and Berger (1994) and Daniels and Kass (2001) consider a reference prior and a set of hierarchical priors respectively that yield posterior shrinkage toward a specified structure.

Regularized likelihood methods for the multivariate Gaussian model provide estimators with different types of shrinkage. Sheena and Gupta (2003) propose a constrained maximum likelihood estimator with constraints on the smallest or the largest eigenvalues. By only focusing on only one of the two ends of the eigenspectrum, this resulting estimator does not correct for the overestimation of the largest eigenvalues and underestimation of the small eigenvalues simultaneously and hence does not address the distortion of the entire eigenspectrum – especially in relatively small sample sizes. Moreover, the choice of regularization parameter and performance comparison with some of the more recently proposed high-dimensional covariance estimators needs to be investigated. Boyd et al. (1998) estimate the Gaussian covariance matrix under the positive definite constraint. In order to exploit sparsity using high-dimensional Gaussian graphical models (Lauritzen, 1996), various ℓ_1-regularization techniques for the elements of the precision or the covariance matrix (or some function thereof) have also been studied by several researchers (Banerjee et al., 2006; Bickel and Levina, 2006; El Karoui, 2008; Friedman et al., 2008; Lam and Fan, 2007; Rajaratnam et al., 2008; Rothman et al., 2008; Yuan and Lin, 2007).

1.1 Shrinkage estimators

We briefly review shrinkage estimators. Letting l_i, $i = 1, \ldots, p$, be the eigenvalues of the sample covariance matrix (sample eigenvalues) in nonincreasing order ($l_1 \geq \ldots \geq l_p \geq 0$), we can decompose the sample covariance matrix as

$$S = Q\mathrm{diag}(l_1, \ldots, l_p)Q^T, \tag{1}$$

where $\mathrm{diag}(l_1, \ldots, l_p)$ 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. Shrinkage estimators have the same eigenvectors but transform the eigenvalues:

$$\widehat{\Sigma} = Q\mathrm{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)Q^T. \tag{2}$$

Typically, sample eigenvalues are shrunk to be more centered, so that the transformed eigenvalues $\widehat{\lambda}_i$ are less spread than those of the sample covariance. In many estimators, the eigenvalues are in the same order as those of the sample covariance: $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$.

Many previous covariance matrix estimators rely explicitly or implicitly on the concept of shrinkage of the eigenvalues of the sample covariance. In the linear shrinkage estimator

$$\widehat{\Sigma}_{\mathrm{LW}} = (1 - \alpha)S + \alpha F, \quad 0 \leq \alpha \leq 1 \tag{3}$$

with the target matrix $F = \gamma I$ for some $\gamma > 0$ (Ledoit and Wolf, 2004b; Warton, 2008), the relationship between the sample eigenvalues l_i and the transformed eigenvalues $\widehat{\lambda}_i$ is affine:

$$\widehat{\lambda}_i = (1 - \alpha)l_i + \alpha\gamma.$$

If F does not commute with S, it does not have the form (2). In Stein's estimator (Stein, 1975, 1977, 1986), the transformed eigenvalues $\widehat{\lambda}_i$ are obtained by applying isotonic regression (Lin and Perlman, 1985) to l_i/γ_i, $i = 1, \ldots, p$ with

$$\gamma_i = \frac{1}{n}\left(n - p + 1 + 2l_i \sum_{j \neq i} \frac{1}{l_i - l_j}\right),$$

in order to maintain the nonincreasing order constraint. In the constrained likelihood approach in Sheena and Gupta (2003), depending on the eigenvalue constraints considered, the shrinkage rule is to truncate the eigenvalues smaller than a *given* lower bound L ($\widehat{\lambda}_i = \max\{l_i, L\}$) or truncate the eigenvalues large than a *given* upper bound U ($\widehat{\lambda}_i = \min\{l_i, U\}$).

1.2 Estimation with a condition number constraint

The condition number of a positive definite matrix Σ is defined as

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

where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are the maximum and the minimum eigenvalues of Σ, respectively. In several applications a stable well-conditioned estimate of the covariance matrix is required. In other words, we require

$$\mathrm{cond}(\Sigma) \leq \kappa_{\max}$$

for a given threshold $\kappa_{\max}$. As an example, in mean-variance (MV) portfolio optimization (Luenberger, 1998; Markowitz, 1952), if the covariance is not well conditioned, the optimization process may amplify estimation error present in the mean return estimate (Ledoit and Wolf, 2004a; Michaud, 1989). The reader is also referred to Ledoit and Wolf (2004a,b) for more extensive discussion of the practical importance of estimating a well-conditioned covariance matrix.

The maximum likelihood estimation problem with the condition number constraint can be formulated as

$$\begin{array}{ll} \text{maximize} & l(\Sigma) \\ \text{subject to} & \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma) \leq \kappa_{\max}. \end{array} \tag{4}$$

(An implicit condition is that Σ is symmetric and positive definite.) This problem is an generalization of the problem considered in Sheena and Gupta (2003), where only either lower bound or upper bound is considered.

The covariance estimation problem (4) can be reformulated as a convex optimization problem, and so can be efficiently solved using standard methods such as interior-point methods 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 $p = 45$.

In Section 2, we show that the condition number constrained estimator $\widehat{\Sigma}_{\mathrm{cond}}$ that solves (4) has the shrinkage form in (2) as

$$\widehat{\Sigma}_{\mathrm{cond}} = Q\mathrm{diag}(\widehat{\lambda}_i, \ldots, \widehat{\lambda}_p)Q^T, \tag{5}$$

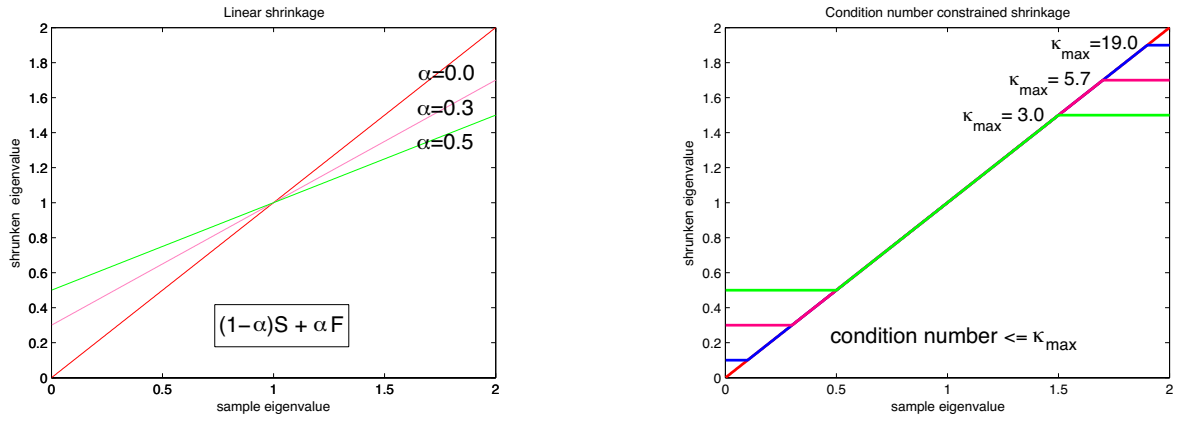


Figure 1: Comparison of eigenvalue shrinkage of the linear shrinkage estimator (left) and the condition number-constrained estimator (right).

with the eigenvalues

$$\widehat{\lambda}_i = \min\big(\max(\tau^\star, l_i), \kappa_{\max}\tau^\star\big), \tag{6}$$

for some $\tau^\star > 0$. In other words, even when the sample size is smaller than the dimension, *i.e.*, $n < p$, the nonlinear shrinkage estimator $\widehat{\Sigma}_{\text{cond}}$ is well-defined. Moreover, the optimal lower bound $\tau^\star$ can be found easily with computational effort $O(p \log p)$, and hence the estimator with the shrinkage rule (6) scales well to much larger size estimation problems, compared with standard solution methods for (4).

The nonlinear transform (6) has a simple interpretation: the eigenvalues of the estimator $\widehat{\Sigma}_{\text{cond}}$ are obtained by truncating the eigenvalues of the sample covariance larger than $\kappa_{\max}\tau^\star$ or smaller than $\tau^\star$. Figure 1 illustrates the transform (6) in comparison with that of the linear shrinkage estimator.

An important issue in the condition number constrained covariance estimation method lies in the selection of $\kappa_{\max}$. We propose a selection procedure that minimizes the predictive risk which is approximated using cross-validation method (Section 3). We show that, for a fixed p, the chosen $\widehat{\kappa}_{\max}$ converges in probability to the condition number of the true covariance matrix as n increases. Furthermore, our numerical study indicates that the selected $\widehat{\kappa}_{\max}$ decreases as p increase. The variance of $\widehat{\kappa}_{\max}$ decreases when either n or p increases.

The risk analysis presented in Section 5 shows that the proposed condition number-constrained estimator $\widehat{\Sigma}_{\text{cond}}$ has smaller risk, with respect to the entropy loss, than the sample covariance matrix S for large n and p. The proposed estimator also performs in general better than other shrinkage estimators in various loss functions. In addition, the condition number-constrained estimator has a smaller condition number than the linear shrinkage estimator particularly when p is large.

1.3 The outline

In the next section, we describe a solution method for the maximum likelihood estimation problem (4). We then propose to estimate the regularization parameter $\kappa_{\max}$ by minimizing predictive risk in Section 3. In Section 4, we give a Bayesian interpretation of the estimator; we show that the prior on the eigenvalues implied by the conditioned number constraint is improper whereas the posterior yields a proper distribution. In Section 5, we compare the risk of the proposed condition number-constrained estimator to those of others, *i.e.*, the sample covariance matrix, Stein's shrinkage estimator, and the linear shrinkage estimator. We prove that the proposed estimator dominates asymptotically the sample covaraince matrix in Stein's risk. Also, we compare numerically the proposed estimator to other estimators in various risks. As an illustrative example, we describe the application in portfolio optimization in Section 6. Finally, we give our conclusions in Section 7. The proofs of the theoretical results discussed in the text are collected in the appendices.

2 Maximum likelihood estimation with the condition number constraint

This section gives the details of the solution (5) and shows how to compute $\tau^\star$. It suffices to consider the case $\kappa_{\max} < l_1/l_p = \mathrm{cond}(S)$, since otherwise the solution to (4) reduces to the sample covariance matrix S.

2.1 Closed-form expression

It is well known that the log-likelihood is a convex function of $\Omega = \Sigma^{-1}$ The condition number constraint on Ω is equivalent to the existence of $u > 0$ such that $uI \preceq \Sigma^{-1} \preceq \kappa_{\max} uI$ where $A \preceq B$ denotes that $B - A$ is positive semidefinite. Since $\mathrm{cond}(\Sigma) = \mathrm{cond}(\Sigma^{-1})$, the covariance estimation problem (4) is equivalent to

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(\Omega S) - \log\det\Omega \\ \text{subject to} & uI \preceq \Omega \preceq \kappa_{\max} uI, \end{array} \tag{7}$$

with variables $\Omega = \Omega^T \in \mathbb{R}^{p\times p}$ and $u > 0$. This problem is a convex optimization problem with $p(p+1)/2 + 1$ variables (Boyd and Vandenberghe, 2004, Chap. 7).

We now show an equivalent formulation with $p+1$ variables. Recall the spectral decomposition of $S = QLQ^T$, with $L = \mathrm{diag}(l_1, \ldots, l_p)$. Suppose the variable Ω has the spectral decomposition $R\Lambda^{-1}R^T$, with R orthogonal and $\Lambda^{-1} = \mathrm{diag}(\mu_1, \ldots, \mu_p)$. Then the objective of (7) is

$$\begin{aligned} \mathbf{Tr}(\Omega S) - \log\det(\Omega) &= \mathbf{Tr}(R\Lambda^{-1}R^T QLQ^T) - \log\det(R\Lambda^{-1}R^T) \\ &= \mathbf{Tr}(\Lambda^{-1}R^T QLQ^T R) - \log\det(\Lambda^{-1}) \\ &\geq \mathbf{Tr}(\Lambda^{-1}L) - \log\det(\Lambda^{-1}). \end{aligned}$$

The equality holds when $R = Q$ (Farrell, 1985, Ch. 14). Therefore we can obtain an equivalent formulation of (7)

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{p}(l_i\mu_i - \log\mu_i) \\ \text{subject to} & u \le \mu_i \le \kappa_{\max} u, \quad i = 1, \ldots, p, \end{array} \tag{8}$$

where the variables are now the eigenvalues $\mu_1, \ldots, \mu_p$ of Λ^{-1}, and u. Let $\mu_1^\star, \ldots, \mu_p^\star, u^\star$ solve (8). The solution to (7) is then

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

We can reduce (8) to an equivalent *univariate* convex problem. We start by observing that (8) is equivalent to

$$\text{minimize} \quad \sum_{i=1}^{p} \min_{u \le \mu_i \le \kappa_{\max} u}(l_i\mu_i - \log\mu_i). \tag{9}$$

Observe that the objective is a separable function of $\mu_1, \ldots, \mu_p$. For a fixed u, the minimizer of each internal term of the objective of (9) is given as

$$\mu_i^\star(u) = \operatorname*{argmin}_{u \le \mu_i \le \kappa_{\max} u} (l_i\mu_i - \log\mu_i) = \min\big\{\max\{u, 1/l_i\}, \kappa_{\max} u\big\}. \tag{10}$$

Then (8) reduces to an unconstrained, univariate optimization problem

$$\text{minimize } J_{\kappa_{\max}}(u), \tag{11}$$

where

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

and

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

This problem is convex, since each $J_{\kappa_{\max}}^{(i)}$ is convex in u. Provided that $\kappa_{\max} < \mathrm{cond}(S)$, (11) has the unique solution

$$u^\star = \frac{\alpha + p - \beta + 1}{\sum_{i=1}^{\alpha} l_i + \sum_{i=\beta}^{p} \kappa_{\max} l_i}, \tag{12}$$

where $\alpha \in \{1, \ldots, p\}$ is the largest index such that $1/l_\alpha < u^\star$ and $\beta \in \{1, \ldots, p\}$ is the smallest index such that $1/l_\beta > \kappa_{\max}u^\star$. Of course both α and β depend on $u^\star$ and cannot be determined *a priori*. However, a simple procedure can find $u^\star$ in $O(p)$ operations on the sample eigenvalues $l_1 \ge \ldots \ge l_p$. This procedure was first considered by Won and Kim (2006) and is elaborated in this paper. The details are given in Appendices A and B.

From the solution $u^\star$ to (11), we can write the solution (5) as

$$\widehat{\Sigma}_{\text{cond}} = Q\text{diag}(\widehat{\lambda}_1^\star, \ldots, \widehat{\lambda}_p^\star)Q^T,$$

where

$$\widehat{\lambda}_i = 1/\mu_i^\star = \min\left\{1/u^\star, \max\{1/(\kappa_{\max}u^\star), l_i\}\right\}$$

solves the covariance estimation problem (4). The eigenvalues of this solution have the form (6), with

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

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

We note that the current univariate optimization method for condition number-constrained estimation is useful for high dimensional problems and is only limited by the complexity of spectral decomposition of the sample covariance matrix (or the singular value decomposition of the data matrix). Our methods is therefore much faster than using interior point methods. We close by noting that this form of estimator is guaranteed to be orthogonally invariant: if the estimator of the true covariance matrix Σ is $\widehat{\Sigma}_{\text{cond}}$, the estimator of the true covariance matrix $U\Sigma U^T$, where U is an orthogonal matrix, is $U\widehat{\Sigma}_{\text{cond}}U^T$.

2.2 A geometric perspective and the regularization path

A simple relaxation of (7) provides an intuitive geometric perspective to the original problem. Consider a function

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

defined as the minimum of the objective of (7) over a fixed range $uI \preceq \Omega \preceq vI$, where $0 < u \leq v$. Following the argument that leads to (9), we can show that

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

Let $\alpha \in \{1, \ldots, p\}$ be the largest index such that $1/l_\alpha < u$ and $\beta \in \{1, \ldots, p\}$ be the smallest index such that $1/l_\beta > v$. Then we can easily show that

$$\begin{aligned} J(u, v) &= \sum_{i=1}^{p} (l_i\mu_i^\star(u, v) - \log\mu_i^\star(u, v)) \\ &= \sum_{i=1}^{\alpha} (l_i u - \log u) + \sum_{i=\alpha+1}^{\beta-1} (1 + \log l_i) + \sum_{i=\beta}^{p} (l_i v - \log v), \end{aligned}$$

where

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

Comparing this to (10), we can observe that $\Omega^\star$ that achieves the minimum in (13) is obtained by truncating the eigenvalues of S greater than $1/u$ and less than $1/v$.

The function $J(u, v)$ has the following properties:

1. J does not increase as u decreases and v increases.
2. $J(u, v) = J(1/l_1, l/l_p)$ for $u \leq 1/l_1$ and $v \geq 1/l_p$. 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 $\{(u, v)|0 < u \leq v\}$, except for on the lines $u = 1/l_1, \ldots, 1/l_p$ and $v = 1/l_1, \ldots, 1/l_p$.

We can now see the following obvious relation between the function $J(u, v)$ and the original problem (7): the solution $u^\star$ to (7) is the minimizer of $J(u, v)$ on the line $v = \kappa_{\max} u$, *i.e.*, $J_{\kappa_{\max}}(u) = J(u, \kappa_{\max} u)$. We denote this minimizer by $u^\star(\kappa_{\max})$.

It would be useful to know how $u^\star(\kappa_{\max})$ behaves as $\kappa_{\max}$ varies. The following result tells us that it has a monotonicity property.

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 almost surely.*

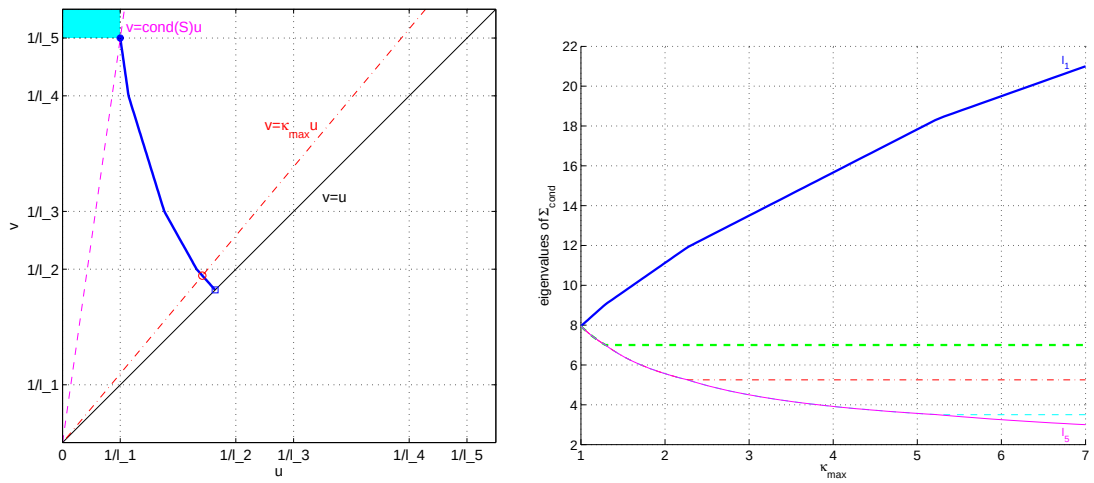
Proof. Given in Appendix C. □

We can plot 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 $(1/l_1, 1/l_p)$ by varying $\kappa_{\max}$ from 1 to $\mathrm{cond}(S)$. Proposition 1 states that, if $\tilde{\kappa}_{\max} > \kappa_{\max}$, the new optimal point $(u^\star(\tilde{\kappa}_{\max}), v^\star(\tilde{\kappa}_{\max}))$ lies on the line segment between the two points:

$$\left(\frac{\kappa_{\max}}{\tilde{\kappa}_{\max}} u^\star(\kappa_{\max}), v^\star(\kappa_{\max})\right), \qquad \left(u^\star(\kappa_{\max}), \frac{\tilde{\kappa}_{\max}}{\kappa_{\max}} v^\star(\kappa_{\max})\right).$$

The proposition also tells us that the optimal truncation range $\left(\tau^\star(\kappa_{\max}), \kappa_{\max}\tau^\star(\kappa_{\max})\right)$ of the sample eigenvalues is nested: once an eigenvalue l_i is truncated for $\kappa_{\max} = \nu_0$, then it keeps truncated for all $\kappa_{\max} < \nu_0$. Hence we have quite a concrete idea of the regularization path of the sample eigenvalues.

Figure 2 illustrates the procedure described above. 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 covariance has eigenvalues $(21, 7, 5.25, 3.5, 3)$. 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 $v = \mathrm{cond}(S)u$ passes ($\kappa_{\max} = \mathrm{cond}(S)$, solid circle). Note that the starting point corresponds to $\widehat{\Sigma}_{\mathrm{cond}} = I$ and the end point to $\widehat{\Sigma}_{\mathrm{cond}} = S$. When $\kappa_{\max} > \mathrm{cond}(S)$, multiple values of $u^\star$ are achieved in the shaded region above the dashed line, nevertheless yielding 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 the truncation ranges of the eigenvalues are nested.



(a) (b)

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 $\kappa_{\max}$.

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

We have discussed so far 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 $\kappa_{\max}$. We describe in this section a criterion for selecting $\kappa_{\max}$.

3.1 Predictive risk selection procedure

We propose to select $\kappa_{\max}$ that minimizes the *predictive risk*, or expected negative predictive log-likelihood

$$\mathrm{PR}(\nu) = \mathbf{E}\left[\mathbf{E}_{\tilde{X}}\left\{ \mathbf{Tr}(\widehat{\Sigma}_\nu^{-1}\tilde{X}\tilde{X}^T) - \log\det\widehat{\Sigma}_\nu^{-1} \right\}\right], \tag{14}$$

where $\widehat{\Sigma}_\nu$ is the estimated condition number-constrained covariance matrix given independent observations $x_1, \cdots, x_n$ from a zero-mean Gaussian distribution on $\mathbb{R}^p$, with the parameter $\kappa_{\max}$ set to ν, and $\tilde{X} \in \mathbb{R}^p$ is a random vector, independent of the given observations, from the same distribution. We approximate the predictive risk using K-fold cross validation. The K-fold cross validation divides the data matrix $\mathbf{X} = (x_1^T, \cdots, x_n^T)$ into K groups so that $\mathbf{X}^T = \left(X_1^T, \ldots, X_K^T\right)$ with n_k observations in the k-th group. For the k-th iteration, each observation in the k-th group X_k plays the role of $\tilde{X}$ in (14), and the remaining $K-1$ groups are used together to estimate the covariance matrix, denoted by $\widehat{\Sigma}_\nu^{[-k]}$. The approximation

of the predictive risk using the k-th group reduces to the predictive log-likelihood

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

The estimate of the predictive risk is then defined as

$$\widehat{\mathrm{PR}}(\nu) = -\frac{1}{n}\sum_{k=1}^{K} l_k\big(\widehat{\Sigma}_\nu^{[-k]}, X_k\big). \tag{15}$$

As the shrinkage parameter $\kappa_{\max}$, we select ν that minimizes (15),

$$\widehat{\kappa}_{\max} = \inf\left\{\nu |\ \mathrm{argmin}_\nu \widehat{\mathrm{PR}}(\nu)\right\}.$$

Note that $l_k\big(\widehat{\Sigma}_\nu^{[-k]}, X_k\big)$ is constant for $\nu \geq \mathrm{cond}(S^{[-k]})$, where $S^{[-k]}$ is the k-th fold sample covariance matrix based on the remaining $q-1$ groups, justifying the use of the smallest minimizer.

3.2 Properties of the selection procedure

It is natural to expect that the estimator $\widehat{\kappa}_{\max}$ has the following properties:

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

(P2) If the true covariance matrix is has a finite condition number, then for given n, $\widehat{\kappa}_{\max}$ approaches to 1 as p increases.

(P3) $\widehat{\kappa}_{\max}$ decreases as p increases.

(P4) The variance of $\widehat{\kappa}_{\max}$ decreases as either n or p increases.

These properties are compatible with the properties of the optimal shrinkage parameter $\widehat{\alpha}$ of the linear shrinkage estimator found using the same predictive risk criterion (Warton, 2008). The difference is that $\widehat{\kappa}_{\max}$ shrinks the sample eigenvalues non-linearly whereas $\widehat{\alpha}$ does linearly.

Because the proposed selection procedure is based on minimizing a numerical approximation of the predictive risk, it is not straightforward to formally validate all the properties given above. At least for (P1), we are able to do so.

Theorem 1. *The estimator $\widehat{\kappa}_{\max}$ satisfies that, for given p,*

$$\lim_{n\to\infty} P\left(\widehat{\kappa}_{\max} = \kappa\right) = 1.$$

Proof. The proof is given in Appendix D. □

As in the case of linear shrinkage estimator (Warton, 2008), we resort to numerical methods to demonstrate (P2)–(P4). To this end, we use data sets sampled from multivariate zero-mean Gaussian distributions with the following covariances:

(i) Identity matrix in $\mathbb{R}^p$.

(ii) $\text{diag}(1, r, r^2, \ldots, r^p)$, with condition number $1/r^p = 5$.

(iii) $\text{diag}(1, r, r^2, \ldots, r^p)$, 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 all combinations of $n \in \{20, 80, 320\}$ and $p \in \{5, 20, 80\}$. For each of these cases, we generate 100 replicates and compute $\widehat{\kappa}_{\max}$ with 5-fold cross validation. The results, plotted in Figure 3, indeed show that the selection procedure described above satisfy the properties (P1)–(P4).

4 Bayesian interpretation

Tibshirani(1996) gives a Bayes interpretation for the Lasso and points outs that in the regression setting, the Lasso solution is equivalent to obtaining the posterior mode when putting a double-exponential (Laplace) prior on the regression coefficients. In the regression setting, the double exponential prior puts relatively more weight near zero (compared to a normal prior)–this Bayesian interpretation gives another perspective on the tendency of the Lasso to set some coefficients to zero and therefore introducing sparsity. In the same spirit, we can draw parallels for the condition number constrained estimator in the covariance estimation problem.

The condition number constraint given by $\lambda_1(\Sigma)/\lambda_p(\Sigma) \leq \kappa_{\max}$ is equivalent to adding a penalty term $g_{\max}\lambda_1(\Sigma)/\lambda_p(\Sigma)$ to the likelihood equation for the eigenvalues. The condition number constrained covariance estimation problem can therefore be written as

$$\begin{array}{ll} \text{maximize} & -\mathbf{Tr}(\Lambda^{-1}L) + \log\det(\Lambda^{-1}) - g_{\max}\,\frac{\lambda_1}{\lambda_p} \\ \text{subject to} & \lambda_1 \geq \lambda_2 \geq \ldots \lambda_p > 0, \end{array}$$

or equivalently, we can write the above maximization problem in terms of the likelihood of the eigenvalues and the penalty as

$$\begin{array}{ll} \text{maximize} & \exp\left(-\frac{n}{2}\sum_{i=1}^{p}\frac{l_i}{\lambda_i}\right)\left(\prod_{i=1}^{p}\lambda_i\right)^{-\frac{n}{2}}\mathrm{e}^{-g_{\max}\left(\frac{\lambda_1}{\lambda_p}\right)} \\ \text{subject to} & \lambda_1 \geq \lambda_2 \geq \ldots \lambda_p > 0 \end{array}$$

The above expression allows us to see the condition number constrained estimator as the Bayes posterior mode under the following prior

$$\pi(\lambda_1, \lambda_1, \ldots, \lambda_p) = \mathrm{e}^{-g_{\max}\left(\frac{\lambda_1}{\lambda_p}\right)}, \qquad \lambda_1 \geq \ldots \geq \lambda_p > 0 \tag{16}$$

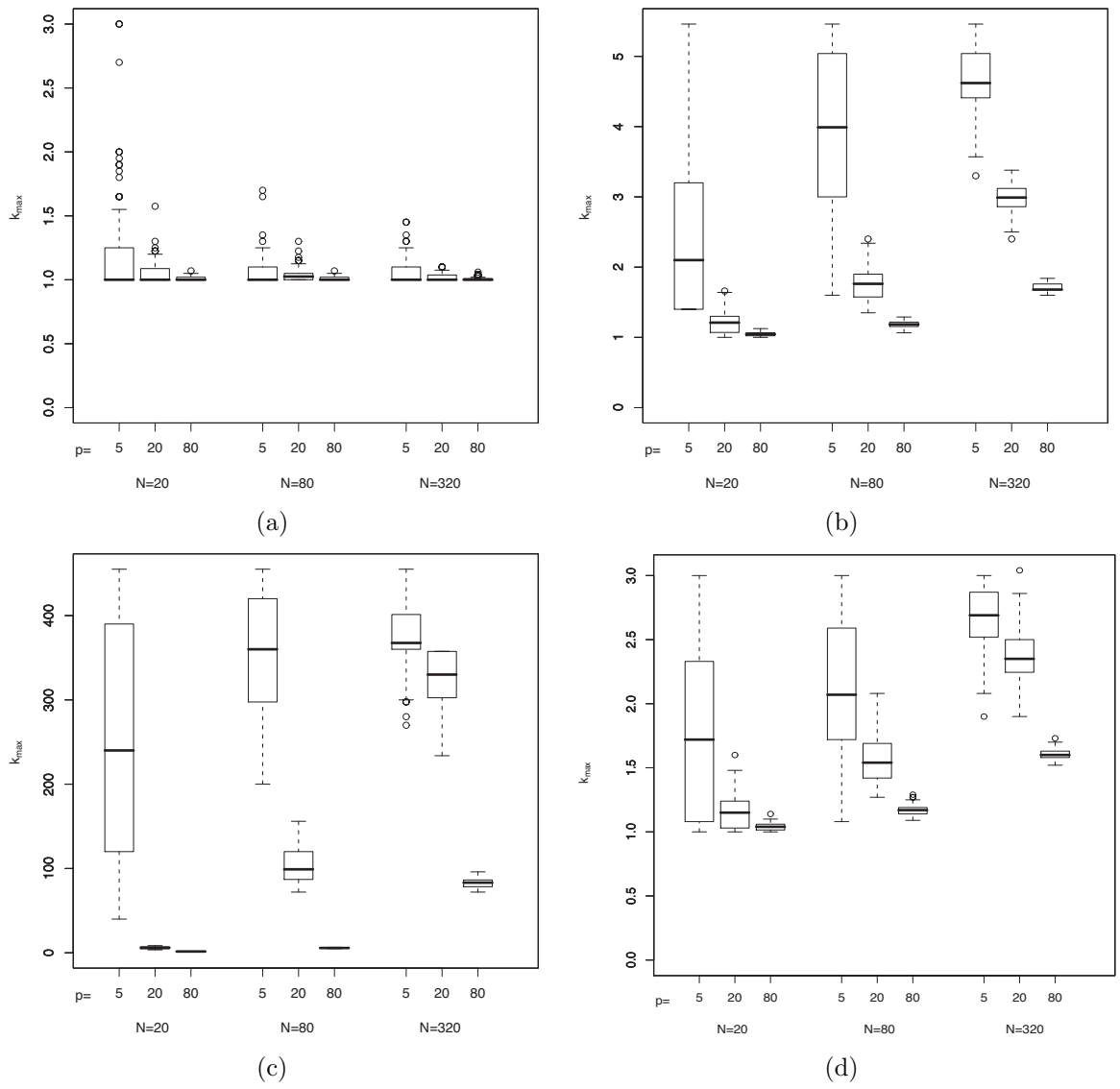


Figure 3: Distribution of $\widehat{\kappa}_{\max}$ for the dimensions 5, 20, 80, and for the sample sizes 20, 80, 320, with covariance matrices (a) identity (b) diagonal exponentially decreasing, condition number 5, (c) diagonal exponentially decreasing, condition number 400, (d) Toeplitz matrix whose (i, j)th element is $0.3^{|i-j|}$ for $i, j = 1, 2, \ldots, p$.

for the eigenvalues and an independent Haar measure on the Stiefel manifold as the prior for the eigenvectors. The prior on the eigenvalues has certain interesting properties which help to explain the type of "truncation" of the eigenvalues that is given by the condition number constrained estimator. First the prior is improper and therefore has an objective or non-informative attribute but always yields a proper posterior:

Proposition 2. *The prior on the eigenvalues implied by the conditioned number constraint is improper whereas the posterior yields a proper distribution. More formally,*

$$\int_C \pi(\underset{\sim}{\lambda}) d\underset{\sim}{\lambda} = \int_C \mathrm{e}^{-g_{\max} \frac{\lambda_1}{\lambda_p}} d\underset{\sim}{\lambda} = \infty,$$

and

$$\int_C \pi(\underset{\sim}{\lambda}) f(\underset{\sim}{\lambda}, \underset{\sim}{l}) d\underset{\sim}{\lambda} \propto \int_C \exp\left(-\frac{n}{2} \sum_{i=1}^{p} \frac{l_i}{\lambda_i}\right) \left(\prod_{i=1}^{p} \lambda_i\right)^{-\frac{n}{2}} \mathrm{e}^{-g_{\max}\left(\frac{\lambda_1}{\lambda_p}\right)} d\underset{\sim}{\lambda} < \infty,$$

where

$$C = \left\{ \underset{\sim}{\lambda} : \lambda_1 \geq \cdots \geq \lambda_p > 0 \right\}.$$

Proof. See Appendix F. □

The prior above also puts the greatest mass around the region $\left\{ \underset{\sim}{\lambda} \in \mathbb{R}^p : \lambda_1 = \cdots = \lambda_p \right\}$ which consequently encourages shrinking or pulling the eigenvalues closer together (see Figure 4).

A clear picture of the type of shrinkage given by the prior above and its potential for "eigenvalue clustering" emerges when compared to the other types of priors suggested in the literature and the corresponding Bayes estimators. The standard MLE of course implies a completely flat prior on the constrained space $C = \left\{ \underset{\sim}{\lambda} : \lambda_1 \geq \cdots \geq \lambda_p > 0 \right\}$. A commonly used prior for covariance matrices is the conjugate prior as given by the inverse-Wishart distribution. The scale hyper-parameter is often chosen to be a multiple of the identity, *i.e.*, $\Sigma^{-1} \sim \text{Wishart}(m, cI)$. This prior yields a posterior mode which is a weighted average of the sample covariance matrix and the prior scale parameter,

$$\Sigma^{\text{post}} = \frac{n}{n+m} S + \frac{m}{n+m} cI.$$

If a prior mode of $\bar{l}I$ is used the posterior mode yields eigenvalue estimates as given by

$$\widehat{\lambda}_i = \frac{n}{n+m} l_i + \frac{m}{n+m} \bar{l}$$

which is simply a linear combination of the sample eigenvalues l_i and the overall mean of the sample eigenvalues as given by $\bar{l}$. Note however that the coefficients of the combination do not depend of the data X and only on the sample size n and m, the degrees of freedom or shape parameter from the prior. This estimator however does not guarantee a well-conditioned estimator for Σ. Ledoit and Wolf (2004) propose an estimator which is also a

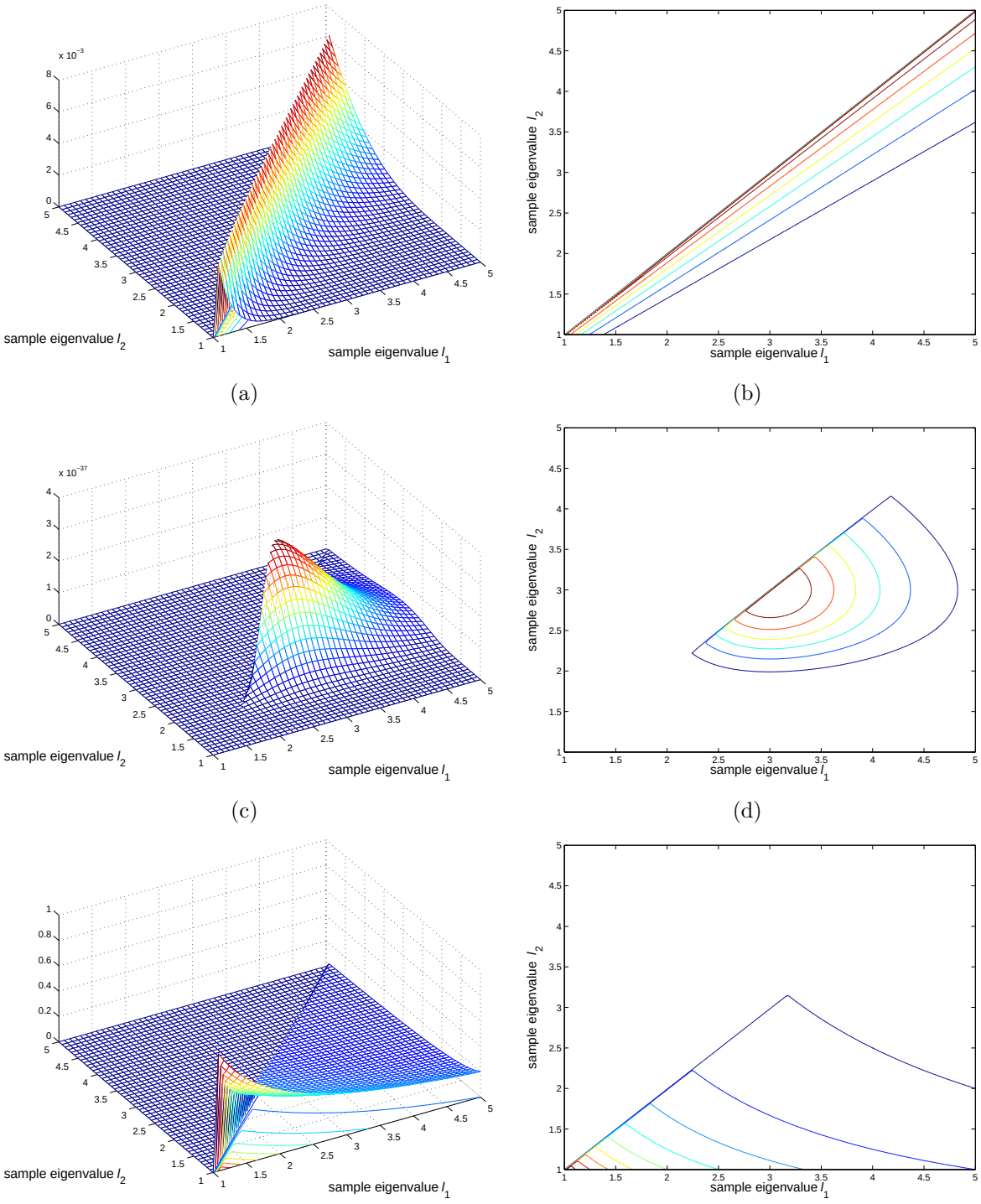


Figure 4: Comparison of various prior densities for sample eigenvalues ($p = 2$). (a) 3-dimensional, (b) contour view of the prior density (16). (c) 3-dimensional, (d) contour view of the prior density induced by the inverse Wishart distribution. (a) 3-dimensional, (b) contour view of the prior density (17) due to Yang and Berger (1994).

linear combination of the sample covariance matrix S and a multiple of the identity but where the coefficients of the combination depends on the data. The Ledoit-Wolf estimator however yields an estimator which is well-conditioned even when $p > n$ and therefore provides a useful tool for estimating Σ in high-dimensional settings - though this estimator is more difficult to interpret as a Bayesian posterior mode.

Yet another useful prior for covariance matrices is the reference prior proposed by Yang and Berger (1994). This prior places an independent Haar distribution on the eigenvectors and the inverse of the Vandermonde determinant as a prior for the eigenvalues and is given by

$$\pi(\lambda_1, \lambda_1, \ldots, \lambda_p, H) = \frac{1}{\prod_{i=1}^{p} \lambda_i}(d\lambda)(dH), \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0, \quad H \in \mathcal{O}(\mathrm{p}). \quad (17)$$

The Vandermonde determinant in the denominator encourages shrinkages of the eigenvalues - though like the conjugate prior and unlike the Ledoit-Wolf estimator, the motivation for the reference prior does not stem from obtaining well-conditioned estimators in high-dimensional problems.

More precisely, simple calculations show that the posterior mode using this reference prior can be formulated as

$$\begin{aligned} &\mathrm{argmax}_{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0} \quad \exp\left(-\frac{n}{2}\sum_{i=1}^{p}\frac{l_i}{\lambda_i}\right)\left(\prod_{i=1}^{p}\lambda_i\right)^{-\frac{n}{2}} \frac{1}{\prod_{i=1}^{p}\lambda_i} \\ = \quad &\mathrm{argmin}_{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0} \quad \frac{n}{2}\sum_{i=1}^{p}\frac{l_i}{\lambda_i} + \frac{n+2}{2}\sum_{i=1}^{p}\log\lambda_i. \end{aligned}$$

An examination of the penalty implied by the reference prior suggests that there is no direct penalty on the condition number. Figure 4 gives the density of the priors discussed above in the two-dimensional case. In particular, the density of the "condition number constraint" prior places more emphasis on the line $\lambda_1 = \lambda_2$ thus "squeezing" the eigenvalues together. This is in direct contrast with the inverse gamma or reference priors where this effect is not as severe.

5 Risk comparison

5.1 Framework for comparing estimators

We compare the risks of the condition number-constrained estimator ($\widehat{\Sigma}_{\mathrm{cond}}$) to that of other estimators in the literature, *i.e.*, the sample covariance matrix (S), the Ledoit-Wolf optimal linear shrinkage estimator ($\widehat{\Sigma}_{\mathrm{LW}}$) and Stein's shrinkage estimator ($\widehat{\Sigma}_{\mathrm{Stein}}$).

We consider four different loss functions in the risk comparisons.

1. The entropy loss, also known as Stein's loss function:

$$\mathcal{L}_{\mathrm{ent}}(\widehat{\Sigma}, \Sigma) = \mathbf{Tr}(\Sigma^{-1}\widehat{\Sigma}) - \log\det(\Sigma^{-1}\widehat{\Sigma}) - p.$$

2. The (squared) Frobenius loss:

$$\mathcal{L}_{\mathrm{F}}(\widehat{\Sigma}, \Sigma) = \left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{F}}^2 = \sum_{ij} \left(\widehat{\sigma}_{ij} - \sigma_{ij}\right)^2,$$

where $\widehat{\sigma}_{ij}$ and σ_{ij} is the (i,j)th element of $\widehat{\Sigma}$ and Σ, respectively.

3. The (negative) predictive likelihood:

$$\mathcal{L}_{\mathrm{P}}(\widehat{\Sigma}, \Sigma) = \mathbf{Tr}(\widehat{\Sigma}^{-1}\Sigma) + \log\det(\widehat{\Sigma}^{-1}) = \mathcal{L}_{\mathrm{ent}}(\widehat{\Sigma}^{-1}, \Sigma^{-1}) + C,$$

where C is a constant which depends on neither $\widehat{\Sigma}$ nor Σ.

4. The quadratic loss:

$$\mathcal{L}_{\mathrm{Q}}(\widehat{\Sigma}, \Sigma) = \left\|\widehat{\Sigma}\Sigma^{-1} - I\right\|_{\mathrm{F}}^2.$$

The risk $\mathcal{R}_i$ of the estimator $\widehat{\Sigma}$ given a loss function $\mathcal{L}_i$ is defined by the expected loss

$$\mathcal{R}_i\big(\widehat{\Sigma}\big) = \mathbf{E}\left[\mathcal{L}_i(\widehat{\Sigma}, \Sigma)\right],$$

where $i = \mathrm{ent}, \mathrm{F}, \mathrm{P}, \mathrm{Q}$.

5.2 Asymptotic dominance

We now show that the condition number constrained estimator $\widehat{\Sigma}_{\mathrm{cond}}$ has asymptotically lower risk with respect to the entropy loss than the sample covariance matrix S. Recall that $\lambda_1, \ldots, \lambda_p$, with $\lambda_1 \geq \cdots \geq \lambda_p$, are the eigenvalues of the true covariance matrix Σ and $\Lambda = \mathrm{diag}\big(\lambda_1, \ldots, \lambda_p\big)$. We further define $\underline{\lambda} = \big(\lambda_1, \ldots, \lambda_p\big)$, $\underline{\lambda}^{-1} = \big(\lambda_1^{-1}, \ldots, \lambda_p^{-1}\big)$, and $\kappa = \lambda_1/\lambda_p$.

We first consider a trivial case in which p/n converges to $\gamma \geq 1$. In this case, the sample covariance matrix S is singular regardless of Σ is singular or not, and $\mathcal{L}_{\mathrm{ent}}(S, \Sigma) = \infty$, whereas both the loss and risk of $\widehat{\Sigma}_{\mathrm{cond}}$ are finite. Thus, $\widehat{\Sigma}_{\mathrm{cond}}$ has smaller entropy risk than S.

We now state the theorem for the case $\gamma < 1$, which shows that, for a properly chosen $\kappa_{\max}$, the condition number-constrained covariance estimator $\widehat{\Sigma}_{\mathrm{cond}} = \widehat{\Sigma}(u^\star)$ of a true covariance matrix with a finite condition number dominates the sample covariance matrix with probability 1.

Theorem 2. *Consider a collection of covariance matrices whose condition numbers are bounded above by κ and whose smallest eigenvalue is bounded below by $u > 0$:*

$$\mathcal{D}(\kappa, u) = \big\{\Sigma = R\Lambda R^T : \ R \text{ othogonal}, \ \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p), u \leq \lambda_1 \leq \ldots \leq \lambda_1 \leq \kappa u\big\}.$$

Then, the following results hold.

(i) For a true covariance matrix Σ, if $\Sigma \in \mathcal{D}(\kappa_{\max}, u)$, then $\widehat{\Sigma}(u)$, which solves (7) for a given u, has a smaller risk with respect to the entropy loss than the sample covariance matrix S.

(ii) For a true covariance matrix Σ whose condition number is bounded above by κ, if $\kappa_{\max} \geq \kappa\big(1-\sqrt{\gamma}\big)^{-2}$, then as $p/n \to \gamma \in (0,1)$,

$$P\left(u^\star \in \left\{u : \Sigma \in \mathcal{D}\big(\kappa_{\max}, u\big)\right\} \ \ \textit{eventually}\ \right) = 1,$$

where $u^\star$ is the solution to (11).

Proof. The proof is given in Appendix E. □

5.3 Monte Carlo comparison

This section conducts a Monte Carlo study to compare the risks of $\widehat{\Sigma}_{\text{con}}$ to those of S, $\widehat{\Sigma}_{\text{LW}}$ and $\widehat{\Sigma}_{\text{Stein}}$. It is worth noting that both S and $\widehat{\Sigma}_{\text{Stein}}$ are not well-defined for $n < p$; their risks are not available.

We use the same simulation setting with that in Section 3. Here we only consider dimensions $p = 20$ and 80. The condition numbers for case (iv) are 3.49 (p=20) and 3.44 (p=80). For each of the simulation scenarios, we generate 100 data sets and compute 100 estimates of the true covariance matrix. We then approximate the risks by taking the average of losses defined in Section 5.1 over 100 estimates.

Table 1–4 summarize risks of the four estimators for cases (i)-(iv), respectively. For case (i) (the true covariance matrix is the identity matrix), $\widehat{\Sigma}_{\text{cond}}$ performs better or as well as $\widehat{\Sigma}_{\text{LW}}$. In particular, when n is smaller than p, $\widehat{\Sigma}_{\text{cond}}$ has smaller condition number than $\widehat{\Sigma}_{\text{LW}}$, and performs better in risk. They both perform much better than S and $\widehat{\Sigma}_{\text{Stein}}$, even when these estimators are defined.

For cases (ii) and (iv) (the true covariance matrices are well-conditioned even though not the identity), $\widehat{\Sigma}_{\text{cond}}$ and $\widehat{\Sigma}_{\text{LW}}$ perform similarly. When defined, $\widehat{\Sigma}_{\text{Stein}}$ performs comparably as well.

For case (iii) (the true covariance matrix is ill-conditioned), $\widehat{\Sigma}_{\text{cond}}$ performs better than $\widehat{\Sigma}_{\text{LW}}$ in the quadratic risk but not in the Frobenius risk. It is also interesting to see that both S and $\widehat{\Sigma}_{\text{Stein}}$ performs as well as $\widehat{\Sigma}_{\text{cond}}$ and $\widehat{\Sigma}_{\text{LW}}$ when n is much greater than p, where sometimes $\widehat{\Sigma}_{\text{cond}}$ deteriorates.

Overall, both $\widehat{\Sigma}_{\text{cond}}$ and $\widehat{\Sigma}_{\text{LW}}$ outperforms S and $\widehat{\Sigma}_{\text{Stein}}$. Although neither $\widehat{\Sigma}_{\text{cond}}$ nor $\widehat{\Sigma}_{\text{LW}}$ dominates each other, $\widehat{\Sigma}_{\text{cond}}$ shows a consistently good risk performance for all the simulation scenarios considered.

6 Application to portfolio selection

The proposed estimator can be useful in a variety of applications that require a covariance matrix estimate that is not only invertible but also well-conditioned. We illustrate the mer-

Table 1: Monte Carlo risk estimation results (case (i))

		$p = 20$				$p = 80$			
n	Loss	Sample	LW	Stein	Condi	Sample	LW	Stein	Condi
20	Entropy	–	0.1546	–	0.0856	–	0.4187	–	0.0524
	Predictive	–	0.1505	–	0.0912	–	0.3691	–	0.0525
	Quadratic	–	0.3298	–	0.1630	–	0.9719	–	0.1047
	Frobenius	–	0.3298	–	0.1623	–	0.9719	–	0.1047
80	Entropy	2.9540	0.0303	0.4540	0.0244	–	6.7636	–	6.6670
	Predictive	4.3557	0.0298	0.5871	0.0247	–	6.3491	–	6.9474
	Quadratic	5.4209	0.0622	0.7310	0.0484	–	16.695	–	14.605
	Frobenius	5.4209	0.0622	0.7310	0.0484	–	13.207	–	14.639
320	Entropy	0.6767	0.0061	0.1098	0.0065	11.093	0.0080	1.6407	0.0070
	Predictive	0.7457	0.0061	0.1215	0.0066	16.002	0.0080	1.9844	0.0071
	Quadratic	1.3150	0.0122	0.1997	0.0128	20.209	0.0161	2.7411	0.0139
	Frobenius	1.3150	0.0122	0.1997	0.0128	20.209	0.0161	2.7411	0.0139

- n: sample size.
- Columns 3 – 7: estimators. `Sample` – sample covariance matrix. `LW` – Ledoit-Wolf optimal linear shrinkage estimator. `Stein` – Stein's estimator. `Condi`– proposed estimator with the maximum condition number $\kappa_{\max}$ chosen via the selection procedure described in Section 3.
- Numbers: estimated risks and their standard deviation with respect to risk defined in Section 4.1. '–' indicates that the estimator is not defined well.

Table 2: Monte Carlo risk estimation results (case (ii))

		$p = 20$				$p = 80$			
n	Loss	Sample	LW	Stein	Condi	Sample	LW	Stein	Condi
20	Entropy	–	2.4853	–	2.4819	–	10.219	–	10.177
	Predictive	–	2.1153	–	2.2679	–	8.5691	–	8.6660
	Quadratic	–	7.2002	–	6.6412	–	29.435	–	28.296
	Frobenius	–	1.0532	–	1.1072	–	4.2372	–	4.2130
80	Entropy	2.9541	1.5130	1.5000	1.5320	–	8.5876	–	8.5148
	Predictive	4.3557	1.3597	1.9478	1.5750	–	7.3545	–	8.0426
	Quadratic	6.7793	4.0837	3.0530	3.6181	–	23.860	–	21.461
	Frobenius	1.3860	0.6583	0.7713	0.7733	–	3.5184	–	3.9182
320	Entropy	0.6767	0.5619	0.5520	0.5732	11.093	5.4080	5.2077	5.5230
	Predictive	0.7457	0.5396	0.5996	0.6081	16.002	4.8704	6.8733	5.6974
	Quadratic	1.6324	1.4224	1.2752	1.3349	25.031	14.319	9.8696	12.756
	Frobenius	0.3370	0.2646	0.2835	0.3000	5.0259	2.2969	2.8406	2.7943

- Symbols are the same as those in Table 1.

Table 3: Monte Carlo risk estimation results (case (iii))

		$p = 20$				$p = 80$			
n	Loss	Sample	LW	Stein	Condi	Sample	LW	Stein	Condi
20	Entropy	–	16.297	–	8.5059	–	466.44	–	468.70
	Predictive	–	7.9300	–	11.526	–	81.081	–	91.499
	Quadratic	–	206.74	–	48.410	–	13553.3	–	13356.2
	Frobenius	–	26.770	–	101.32	–	3.2404	–	4.4762
80	Entropy	2.9541	3.4144	2.0478	2.3943	–	217.58	–	129.21
	Predictive	4.3557	2.4553	2.8252	3.2353	–	51.413	–	58.772
	Quadratic	46.066	47.423	29.209	21.240	–	3846.7	–	1535.4
	Frobenius	6.5707	6.6223	11.381	57.249	–	1.6247	–	3.4484
320	Entropy	0.6767	0.7116	0.5964	0.6359	11.093	57.583	9.5260	11.441
	Predictive	0.7457	0.6380	0.6559	0.6943	16.002	21.990	12.999	13.937
	Quadratic	11.470	11.557	10.086	8.6066	232.95	579.05	153.19	159.03
	Frobenius	1.5072	1.4844	1.8293	8.3058	0.6033	0.5370	0.7104	1.1705

- Symbols are the same as those in Table 1.

Table 4: Monte Carlo risk estimation results (case (iv))

		$p = 20$				$p = 80$			
n	Loss	Sample	LW	Stein	Condi	Sample	LW	Stein	Condi
20	Entropy	–	1.8262	–	1.7760	–	8.0696	–	7.8528
	Predictive	–	1.7015	–	1.7866	–	7.4562	–	7.4731
	Quadratic	–	4.6845	–	4.0940	–	20.591	–	18.790
	Frobenius	–	3.5973	–	3.6764	–	15.965	–	15.641
80	Entropy	2.9541	1.1599	1.2690	1.1410	–	6.7636	–	6.6670
	Predictive	4.3557	1.1039	1.6601	1.2365	–	6.3491	–	6.9474
	Quadratic	6.3816	2.8523	2.4157	2.4392	–	16.695	–	14.605
	Frobenius	5.4461	2.2933	2.7431	2.6211	–	13.207	–	14.639
320	Entropy	0.6767	0.4929	0.4884	0.4824	11.093	4.4812	4.7044	4.4483
	Predictive	0.7457	0.4828	0.5335	0.5116	16.002	4.2760	6.3861	4.8584
	Quadratic	1.5485	1.1815	1.0721	1.0737	24.112	11.017	8.3825	9.4276
	Frobenius	1.322	0.9889	1.0533	1.0954	20.286	8.8460	11.092	10.514

- Symbols are the same as those in Table 1.

its of the proposed estimator in portfolio optimization. We consider a portfolio rebalancing strategy based on minimum variance portfolio selection, since it relies only on the covariance not on the mean return vector which is known to be extremely difficult to estimate (Luenberger, 1998; Merton, 1980). For this reason, minimum variance portfolio selection has been studied extensively in the literature; see, *e.g.*, Chan et al. (1999).

We use the proposed estimator, combined with the $\kappa_{\max}$ selection procedure described above, and two existing ones, namely, the Ledoit-Wolf optimal linear shrinkage estimator (LW) and the sample covariance, in constructing a minimum variance portfolio. We compare their performance over a period of more than 14 years. It is therefore necessary to rebalance the portfolio to account for possible model changes.

6.1 Minimum variance portfolio rebalancing

We have n risky assets, denoted $1, \ldots, n$, which are held over a period of time. We use r_i to denote the relative price change (return) of asset i over the period, that is, its change in price over the period divided by its price at the beginning of the period. Let Σ denote the covariance of $r = (r_1, \ldots, r_n)$. We let w_i denote the amount of asset i held throughout the period. (A long position in asset i corresponds to $w_i > 0$, and a short position in asset i corresponds to $w_i < 0$.)

In mean-variance optimization, the risk of a portfolio $w = (w_1, \ldots, w_n)$ is measured by the standard deviation $(w^T \Sigma w)^{1/2}$ of its return (Markowitz, 1952). 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 minimum variance portfolio optimization problem can be formulated as

$$\begin{array}{ll} \text{minimize} & w^T \Sigma w \\ \text{subject to} & \mathbf{1}^T w = 1. \end{array}$$

This simple quadratic program has an analytic solution given by

$$w = \frac{1}{\mathbf{1}^T \Sigma^{-1} \mathbf{1}} \Sigma^{-1} \mathbf{1}.$$

The portfolio selection method described above assumes that the returns are stationary, which does not hold in reality. As a way of coping with the nonstationarity of returns, we describe the minimum variance portfolio rebalancing (MVR) strategy. Let $r_t = (r_{1t}, \ldots, r_{nt}) \in \mathbb{R}^n$, $t = 1, \ldots, N_{\text{tot}}$, be the realized returns of assets at time t. (The time unit can be a day, a week, or a month.) We consider periodic minimum variance rebalancing in which the portfolio weights are updated in every L time units. After observing the close prices of the assets at the end of each period, we select the minimum variance portfolio with the data available till the moment and hold it for the next L time units. Let N_{estim} be the estimation horizon size, *i.e.*, the number of past data points used to estimate the covariance. For simplicity, we assume

$$N_{\text{tot}} = N_{\text{estim}} + KL,$$

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 is not taken into account.) We therefore have a series of portfolios

$$w^{(j)} = \frac{1}{\mathbf{1}^T \hat{\Sigma}_j^{-1} \mathbf{1}} \hat{\Sigma}_j^{-1} \mathbf{1}$$

over the periods of $[N_{\text{estim}}+1+(j-1)L, N_{\text{estim}}+jL]$, $j = 1, \ldots, K$. $\hat{\Sigma}_j$ is the covariance of the asset returns estimated from the asset returns over the horizon $[1+(j-1)L, N_{\text{estim}}+(j-1)L]$.

6.2 A numerical example

In our experimental study, we use the 30 stocks that constituted the Dow Jones Industrial Average over the period from February 1994 to July 2008. Table 5 shows the 30 stocks. We used adjusted close prices, namely, the closing prices day adjusted for all applicable splits and dividend distributions, which were downloaded from Yahoo finance (`http://finance.yahoo.com/`). (The data are adjusted using appropriate split and dividend multipliers, in accordance with Center for Research in Security Prices (CRSP) standard.)

The whole period considered in our numerical study is from the first trading date in March 2, 1992 to July 14, 2008. (The whole horizon consists of 4125 trading days.) The time unit used in our numerical study is 5 consecutive trading days, so we consider weekly returns. We take

$$N_{\text{tot}} = 825, \qquad L = 25, \qquad N_{\text{estim}} = 100.$$

To compute an estimate the covariance, we use the last $N_{\text{estim}} = 100$ weekly returns of the constituents of the Dow Jones Industrial Average. Roughly in every half year, we rebalance the portfolio, using a covariance estimate with past roughly two-year weekly return data. Table 6 shows the periods determined by the choice of the parameters. The trading period is from 5 to 33 ($K = 29$), which spans the period from February 18, 1994 to July 14, 2008. For the ith period, we use the data from the beginning of the $i-4$th period to the end of the $i-1$th period to estimate the covariance of the asset returns and hold the corresponding minimum variance portfolio over the ith period.

In the sequel, we compare the MVR strategy where the covariance is estimated by using the proposed estimator with the MVR strategies using the sample covariance and the LW estimator.

Performance metrics

We will use the following quantities in assessing the performance of the MVR strategies.

- *Realized return.* The realized return of a portfolio w over the period $[N_{\text{estim}} + 1 + (j-1)L, N_{\text{estim}} + jL]$ is computed as

$$r_j(w) = \sum_{t=N_{\text{estim}}+1+(j-1)L}^{N_{\text{estim}}+jL} r_{it} w_i.$$

Table 5: Dow Jones stocks used in our numerical study and their market performance over the period from February 18, 1994 to July 14, 2008. The return, risk and SR are annualized.

index	company	ticker	return [%]	risk [%]	SR	MDD [%]
1	3M Company	MMM	12.04	10.74	0.25	32.20
2	Alcoa, Inc.	AA	16.50	15.47	0.30	59.34
3	American Express	AXP	17.52	14.61	0.35	59.07
4	American International Group, Inc.	AIG	7.96	12.93	0.07	77.07
5	AT&T Inc.	T	11.57	12.95	0.19	64.66
6	Bank of America	BAC	11.57	13.14	0.19	59.80
7	The Boeing Company	BA	13.03	13.81	0.23	62.82
8	Caterpillar Inc.	CAT	18.53	14.26	0.39	52.32
9	Chevron Corporation	CVX	15.86	10.53	0.42	33.62
10	Citigroup Inc.	C	14.44	15.27	0.25	71.10
11	The Coca-Cola Company	KO	10.74	10.77	0.20	55.44
12	E.I. du Pont de Nemours & Company	DD	9.58	12.43	0.13	55.70
13	Exxon Mobil Corporation	XOM	16.58	10.46	0.45	33.72
14	General Electric Company	GE	13.47	12.04	0.28	61.87
15	General Motors Corporation	GM	-1.24	15.85	-0.20	85.71
16	The Hewlett-Packard Company	HPQ	20.22	18.24	0.35	82.90
17	The Home Depot	HD	12.96	15.28	0.20	69.74
18	Intel Corporation	INTC	20.84	19.13	0.35	82.24
19	International Business Machines Corp.	IBM	20.99	13.86	0.48	59.36
20	Johnson & Johnson	JNJ	17.13	10.10	0.49	35.89
21	JPMorgan Chase & Co.	JPM	15.84	15.44	0.29	74.02
22	McDonald's Corporation	MCD	14.05	12.05	0.30	73.60
23	Merck & Co., Inc.	MRK	12.86	12.87	0.24	68.00
24	Microsoft Corporation	MSFT	22.91	15.13	0.50	65.15
25	Pfizer Inc.	PFE	15.34	12.92	0.32	57.15
26	The Procter & Gamble Company	PG	15.25	11.06	0.37	54.22
27	United Technologies Corporation	UTX	18.93	12.37	0.47	52.10
28	Verizon Communications Inc.	VZ	9.93	12.38	0.14	56.82
29	Wal-Mart Stores, Inc.	WMT	14.86	13.16	0.30	37.49
30	The Walt Disney Company	DIS	10.08	14.05	0.13	67.90

Table 6: Trading periods

index	period	index	period
1	3/02/1992 – 8/26/1992	2	8/27/1992 – 2/24/1993
3	2/25/1993 – 8/23/1993	4	8/24/1993 – 2/17/1994
5	2/18/1994 – 8/18/1994	6	8/19/1994 – 2/15/1995
7	2/16/1995 – 8/15/1995	8	8/16/1995 – 2/12/1996
9	2/13/1996 – 8/09/1996	10	8/12/1996 – 2/06/1997
11	2/07/1997 – 8/06/1997	12	8/07/1997 – 2/04/1998
13	2/05/1998 – 8/04/1998	14	8/05/1998 – 2/02/1999
15	2/03/1999 – 8/02/1999	16	8/03/1999 – 1/28/2000
17	1/31/2000 – 7/27/2000	18	7/28/2000 – 1/25/2001
19	1/26/2001 – 7/25/2001	20	7/26/2001 – 1/29/2002
21	1/30/2002 – 7/29/2002	22	7/30/2002 – 1/27/2003
23	1/28/2003 – 7/25/2003	24	7/28/2003 – 1/23/2004
25	1/26/2004 – 7/23/2004	26	7/26/2004 – 1/20/2005
27	1/21/2005 – 7/20/2005	28	7/21/2005 – 1/18/2006
29	1/19/2006 – 7/18/2006	30	7/19/2006 – 1/17/2007
31	1/18/2007 – 7/17/2007	32	7/18/2007 – 1/14/2008
33	1/15/2008 – 7/14/2008		

- *Realized risk over the jth period.* The realized risk (return standard deviation) of a portfolio w over the period $[N_{\text{estim}} + 1 + (j-1)L, N_{\text{estim}} + jL]$ is computed as

$$\sigma_j(w) = w^T \Sigma_{\text{sample}}^{(j)} w,$$

  where $\Sigma_{\text{sample}}^{(j)}$ is the sample covariance of the asset returns over the period.
- *Realized Sharpe ratio (SR).* The realized Sharpe ratio, *i.e.*, the ratio of the excess expected return of a portfolio w relative to the risk-free return μ_{rf} is given by

$$S_k(w) = \frac{\sigma_j(w) - \mu_{\text{rf}}}{r_j(w)}.$$

- *Turnover.* The turnover from the portfolio held at the start date of the jth period w^j to the portfolio $w^{(j-1)}$ held at the previous period is computed as

$$\text{TO}(j) = \sum_{i=1}^{n} \left| w_i^{(j)} - \prod_{t=N_{\text{estim}}+1+(j-1)L}^{N_{\text{estim}}+jL} r_{it} w_i^{(j-1)} \right|.$$

  For the first period, we take $w^{(0)} = 0$, *i.e.*, the initial holdings of the assets are zero.

- *Transaction cost.* If the cost to buy or sell one share of stock i is η_i, then the transaction cost due to the rebalancing is

$$\mathrm{TC}(j) = \sum_{i=1}^{n} \eta_i \left| w_i^{(j)} - \prod_{t=N_{\mathrm{estim}}+1+(j-1)L}^{N_{\mathrm{estim}}+jL} r_{it} w_i^{(j-1)} \right|.$$

Let $w^{(j)} = (w_1^{(j)}, \ldots, w_n^{(j)})$ be the portfolio constructed by a rebalancing strategy held over the period $[N_{\mathrm{estim}} + 1 + (j-1)L, N_{\mathrm{estim}} + jL]$. When the initial budge is normalized to one, the normalized wealth grows according to the recursion

$$W(t) = \begin{cases} W(t-1)(1 + \sum_{i=1}^{n} w_{it} r_{it}), & t \notin \{N_{\mathrm{estim}} + jL \mid j = 1, \ldots, K\}, \\ W(t-1)(1 + \sum_{i=1}^{n} w_{it} r_{it}) - \mathrm{TC}(j), & t = N_{\mathrm{estim}} + jL, \end{cases}$$

for $t = N_{\mathrm{estim}}, \ldots, N_{\mathrm{estim}} + KL$, with the initial wealth

$$W(N_{\mathrm{estim}}) = 1.$$

Here

$$w_{it} = \begin{cases} w_i^{(1)}, & t = N_{\mathrm{estim}} + 1, \ldots, N_{\mathrm{estim}} + L, \\ \vdots & \\ w_i^{(K)}, & t = N_{\mathrm{estim}} + 1 + (K-1)L, \ldots, N_{\mathrm{estim}} + KL. \end{cases}$$

Another performance metric of interest is the maximum drawdown, *i.e.*, the maximum cumulative loss from a market peak to the following trough. The maximum drawdown at time t is the largest drawdown of the wealth experienced by the trading strategy up to time t:

$$\mathrm{MDD}(t) = \frac{W(t)}{\sup_{s \le t} W(s)}.$$

The maximum drawdown is the maximum cumulative loss from a peak to the following trough:

$$\mathrm{MD} = \sup_{t = N_{\mathrm{estim}}, \ldots, N_{\mathrm{trading}} + KL} \mathrm{MDD}(t).$$

Comparison results

We assume that the transaction costs are the same for the 30 stocks:

$$\eta_i = \bar{\eta}.$$

We set $\bar{\eta}$ to 30 basis points (bps). Since the rebalancing is done biannually, the wealth growth does not depend very much on transaction costs, so long as they are below 40 bps. We set the risk-free return as $\mu_{\mathrm{rf}} = 0.05$, *i.e.*, 5% per annum.

Figure 5 shows the wealth growth over the trading horizon, from the start date of the 5th period to the final date of the 33th period (*i.e.*, from February 18, 1994 through July 14,

2008). The MVR strategy using the proposed estimator outperforms significantly the MVR strategy using the sample covariance or the LW covariance estimator. Table 7 summarizes their performance. For comparison, Table 8 gives the descriptive statistics of the S&P 500 index over the same period. The MVR strategy using the proposed estimator outperformed significantly the S&P 500 index. The maximum drawdown achieved by the use of the proposed estimator is relatively low, compared with that of the S&P 500 index over the same period (around 50%).

Table 5 summarizes the performance of the individual Dow Jones stocks over the same period. We can see that the SR achieved by the proposed estimator is comparable to that of the best performing stock while achieving a far lower maximum drawdown.

Figure 7 shows how the MVR strategies perform over the entire trading period. We can see that the use of the sample covariance leads to very volatile results. The use of the LW leads to less volatile results. The use of the proposed estimator leads to least volatile results.

Figure 6 shows how the condition number of the covariance estimate varies over the trading horizon. The proposed covariance estimator leads to a better-conditioned covariance estimate. As a result, the MVR strategy using the proposed estimator produce more stable weights, as can be seen from Figure 8–Figure 10 that show the weights generated by the MVR strategies using the three covariance estimators over the trading horizon.

Thus far, the estimation horizon size has been fixed to four years. We vary the parameter from two years to two years, while fixing the last date of the estimation period and the start date of the trading period to the same values. (For instance, when the estimation horizon size is four years, we use the stock data over the period of February 1996 to July 2006.) Tables 9 and 10 summarize the results. In both cases, the MVR using the proposed estimator outperforms the MVR using the sample covariance or the LW estimator.

Figure 11 shows the plot of turnover versus period, and Table 11 shows the annual turnovers of the rebalancing strategies under consideration. This figure shows that MVR using the proposed estimator gives a lower turnover and more stable weights than MVR using the LW or the sample covariance. Furthermore, the MVR using the proposed estimator gives a lower turnover rate than the MVR using the existing estimators. The annual turnover of the MVR using the proposed estimator is far lower than one using the sample covariance or the LW estimator.

In summary, the portfolio rebalancing strategy based on the proposed covariance estimation method outperforms the portfolio rebalancing strategies based on the LW covariance estimation and sample covariance. The main improvement seems to stem from the fact that since the proposed estimation method leads to a better conditioned estimate than the other two existing methods, while estimating well the principal directions in the covariance.

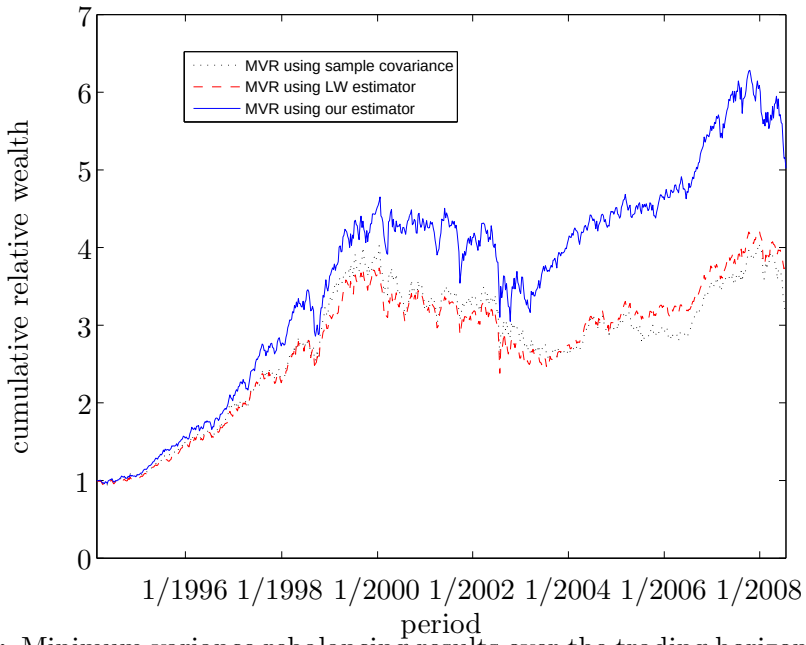


Figure 5: Minimum variance rebalancing results over the trading horizon from February 18, 1994 through July 14, 2008.

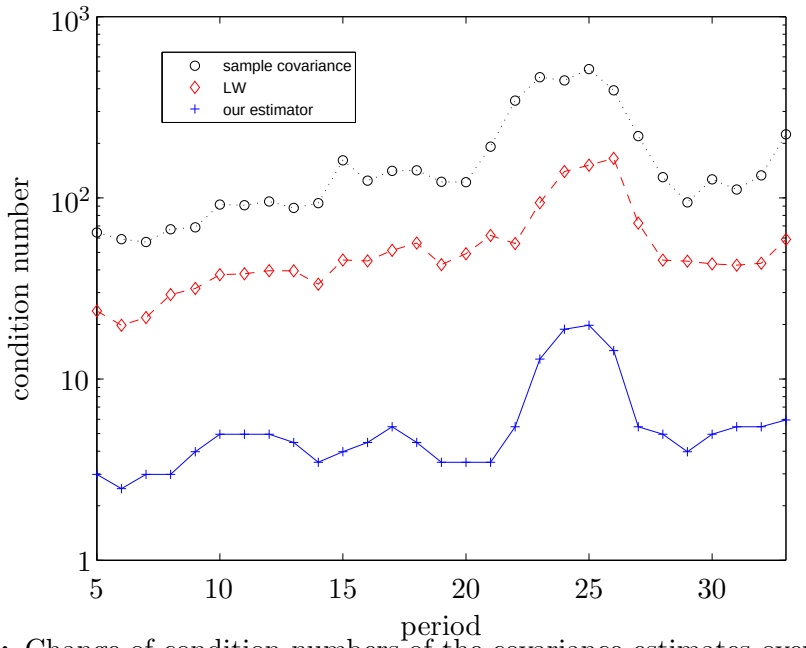


Figure 6: Change of condition numbers of the covariance estimates over the trading horizon from February 18, 1994 through July 14, 2008.

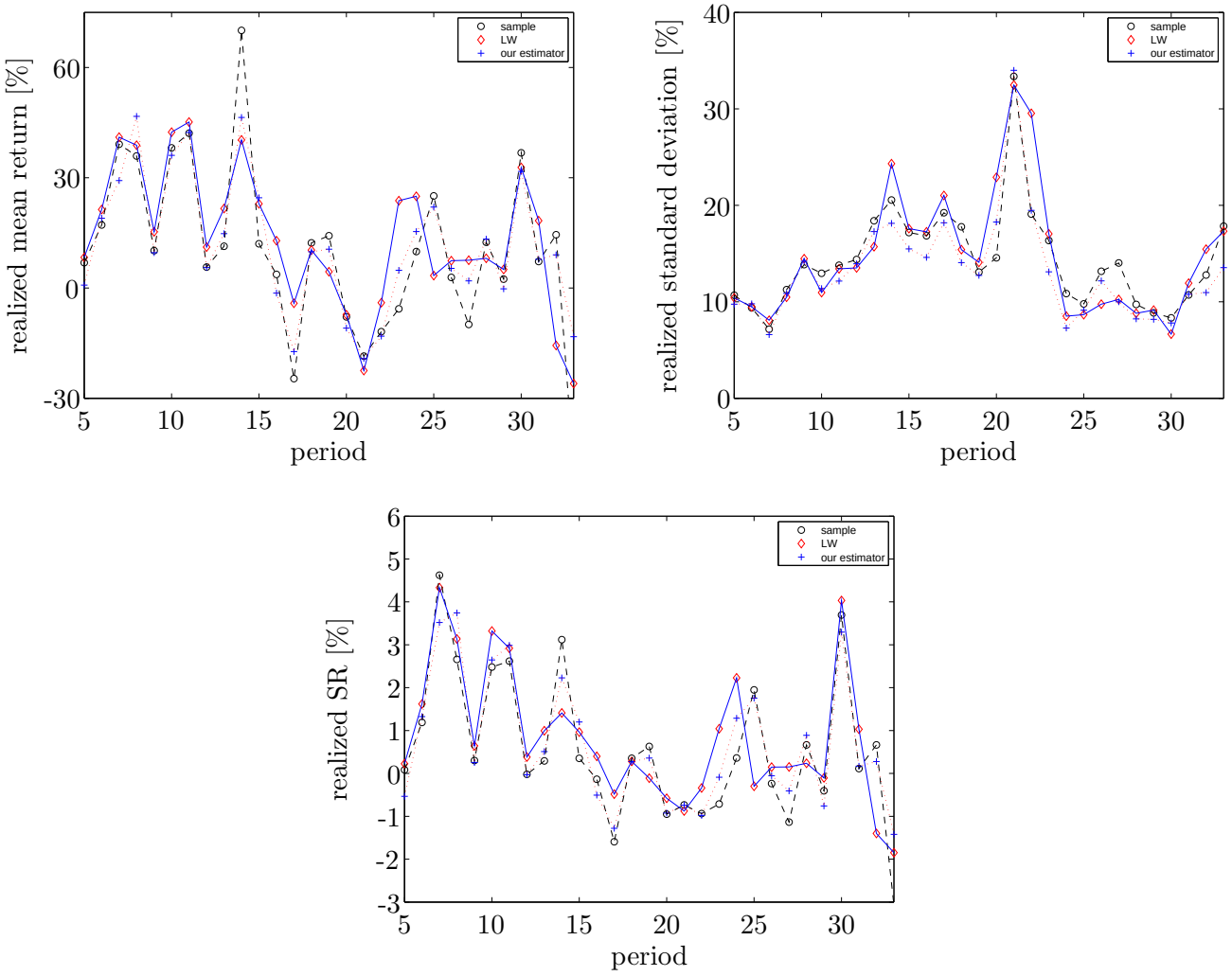


Figure 7: Performance comparison of the three MVR strategies. Top left: realized returns. Top right: realized standard deviations. Bottom: realized Sharpe ratios.

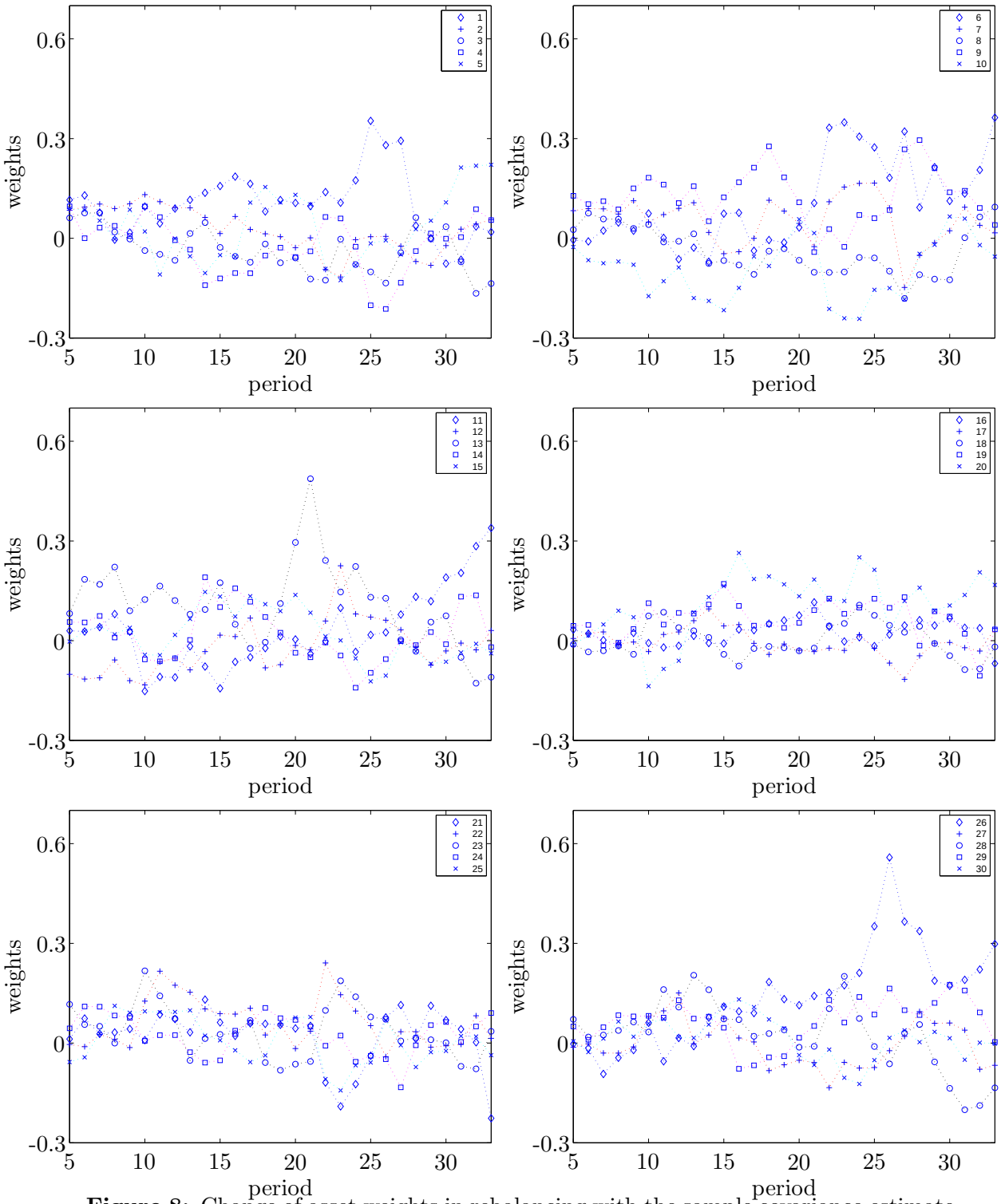


Figure 8: Change of asset weights in rebalancing with the sample covariance estimate from the asset return data over the past period.

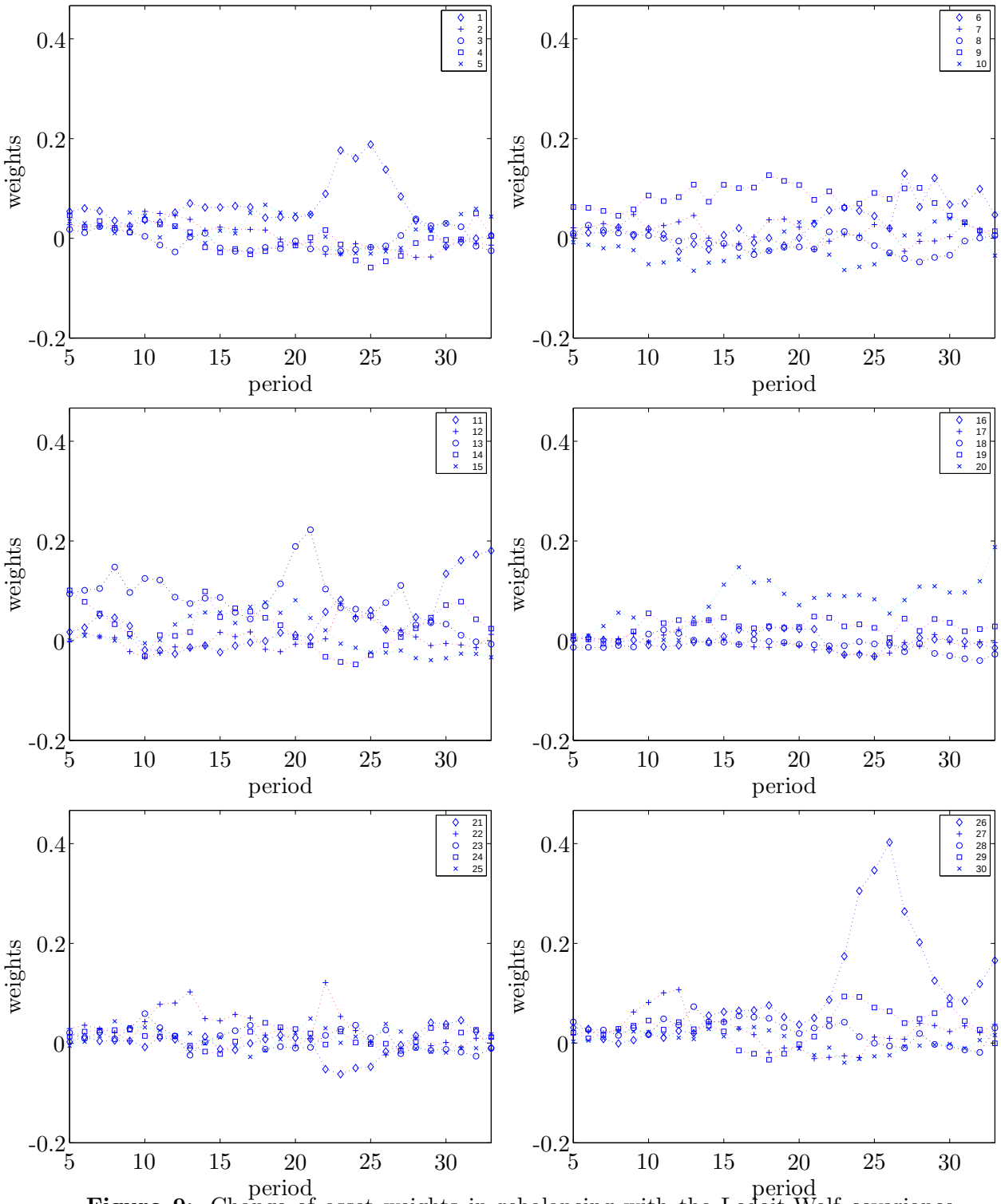


Figure 9: Change of asset weights in rebalancing with the Ledoit-Wolf covariance estimate from the asset return data over the past period

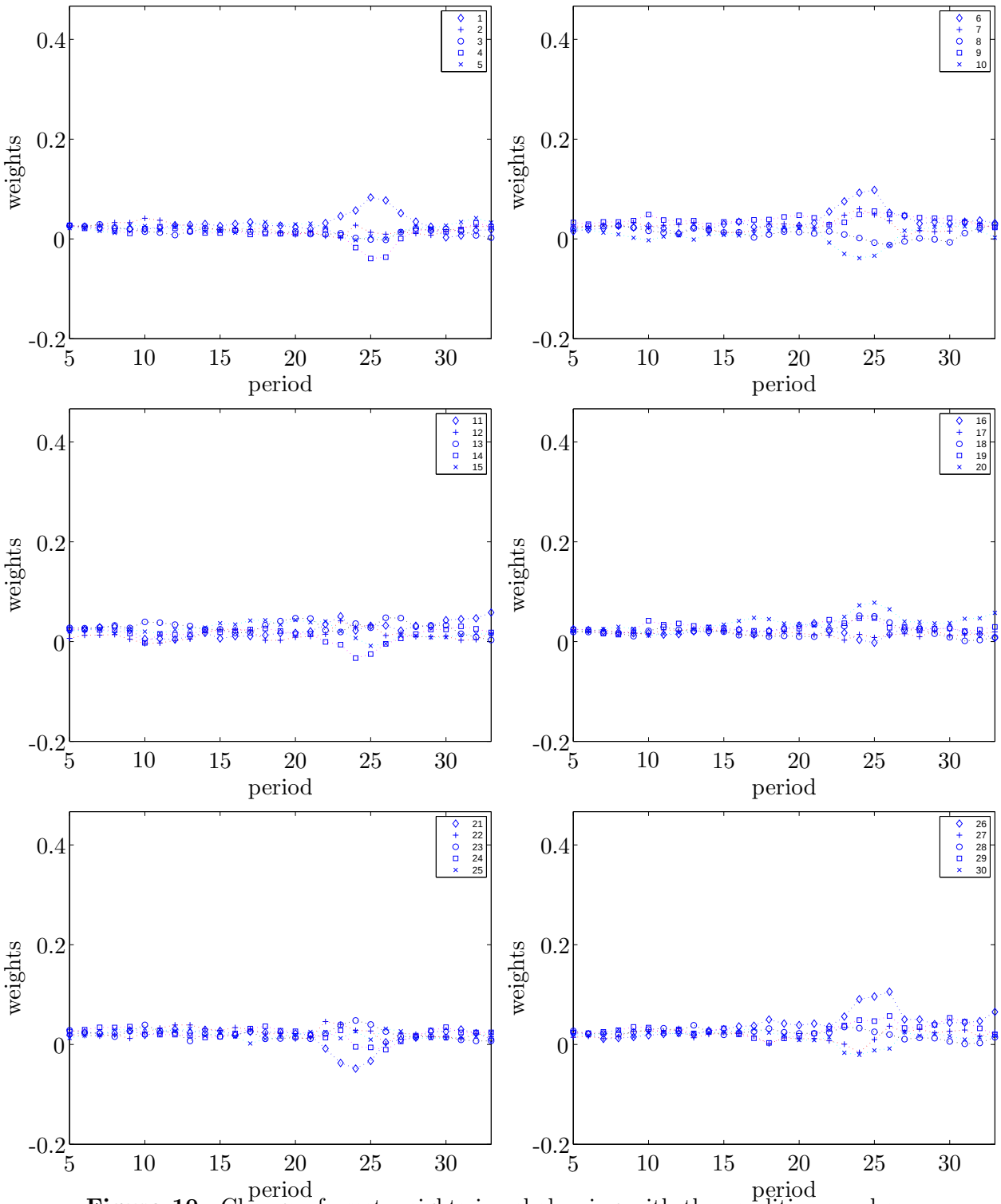


Figure 10: Change of asset weights in rebalancing with the condition-number constrained covariance estimate from the asset return data over the past period when the maximum condition number $\kappa_{\max}$ is chosen via the selection procedure described in Section 3.

Table 7: Performance of the minimum variance portfolio strategy based on different covariance estimation methods over the trading horizon from February 18, 1994 through July 14, 2008.

covariance estimation method	annualized return [%]	annualized risk [%]	annualized SR	maximum drawdown [%]
sample covariance	9.18	15.11	0.21	36.57
LW	10.26	14.05	0.30	36.26
our estimator	12.36	15.54	0.41	34.43

Table 8: Performance of the S&P 500 index over the trading horizon from February 18, 1994 through July 14, 2008.

annualized return [%]	annualized risk [%]	annualized SR	maximum drawdown [%]
7.73	19.36	0.15	49.15

7 Conclusions

In this paper we have considered covariance estimation in the likelihood framework with a constraint on the condition number. Estimators that have been proposed in the literature for high dimensional covariance estimation do not directly target the issue of invertibility and conditioning. We have emphasized the importance of a numerically well-conditioned estimator of covariance matrices, especially in practical applications such as portfolio selection. A consequence of this emphasis on numerical stability is the condition number-constrained maximum likelihood estimator described in this paper. We have shown that this estimator involves optimal truncation of the eigenvalues of the sample covariance matrix. The truncation range is shown to be simple to compute. We have studied how the truncation range varies as a function of the regularization parameter. We have also provided a cross-validated parameter selection procedure. The cross-validated estimator demonstrates a robust risk performance compared with other commonly used estimators of covariance matrices. We study the theoretical properties of our estimator and show that the proposed estimator asymptotically dominates the sample covariance estimator with respect to the entropy loss under a mild assumption. When applied to the real-world wealth management problem, the

Table 9: Performance of the MVR strategies over the trading horizon (from February 16, 1995 through July 14, 2008) when the estimation horizon size is 3 years.

covariance estimation method	annualized return [%]	annualized risk [%]	annualized SR	maximum drawdown [%]
sample covariance	10.65	14.33	0.32	37.18
LW	11.08	14.10	0.36	35.42
our estimator	12.73	15.73	0.43	34.76

Table 10: Performance of the MVR strategies over the trading horizon (from February 13, 1996 through July 14, 2008) when the estimation horizon size is 4 years.

covariance estimation method	annualized return [%]	annualized risk [%]	annualized SR	maximum drawdown [%]
sample covariance	9.18	15.11	0.21	36.57
LW	10.26	14.05	0.31	36.26
our estimator	12.36	15.54	0.41	34.43

Table 11: Average annual turnovers of the rebalancing strategies over the trading horizon from February 18, 1994 through July 14, 2008.

sample covariance	LW	our estimator
303.2%	150.6%	69.0%

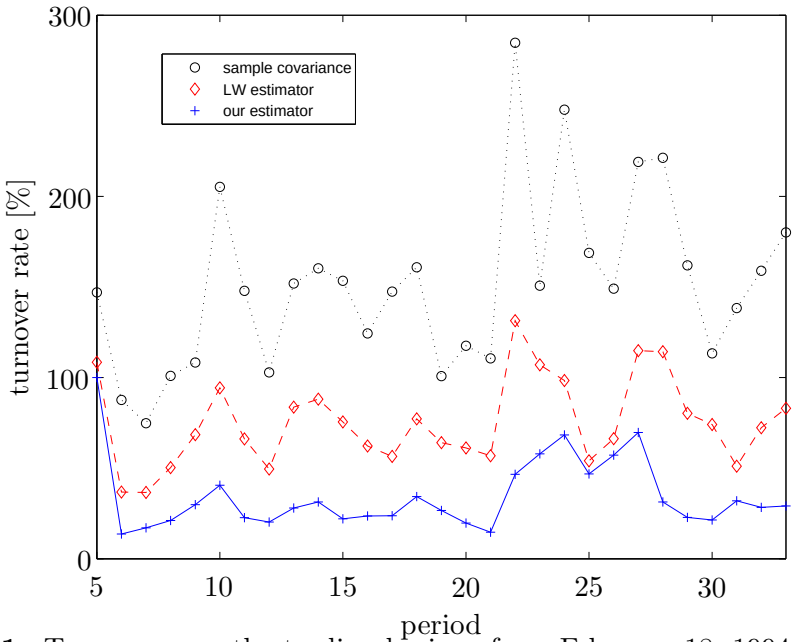


Figure 11: Turnover over the trading horizon from February 18, 1994 through July 14, 2008.

condition number-constrained estimator performs very well, supporting its usefulness in a variety of applications where a well-conditioned covariance estimator is desirable.

Acknowledgments

The authors thank Professors Charles Stein, Richard A. Olshen, Robert M. Gray and Dr. Alessandro Magnani for helpful comments and suggestions.

References

Anderson, T. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In R. Bose (Ed.), *Essays in Probability and Statistics*, pp. 1–24. University of North Carolina Press.

Banerjee, O., A. d'Aspremont, and G. Natsoulis (2006). Convex optimization techniques for fitting sparse Gaussian graphical models. *Proceedings of the 23rd international conference on Machine learning*, 89–96.

Bickel, P. J. and E. Levina (2006). Regularized estimation of large covariance matrices. Technical Report 716, Dept. of Statistics, University of California, Berkeley, CA.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Boyd, S., L. Vandenberghe, and S. P. Wu (1998). Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Applic 19*, 499–533.

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

Daniels, M. and R. Kass (2001). Shrinkage estimators for covariance matrices. *Biometrics 57*, 1173–1184.

Dempster, A. P. (1972). Covariance Selection. *Biometrics 28*(1), 157–175.

Dey, D. K. and C. Srinivasan (1985). Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics 13*(4), 1581–1591.

El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics 36*(6), 2757–2790.

Farrell, R. H. (1985). *Multivariate calculation*. Springer-Verlag New York.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Geman, S. (1980). A limit theorem for the norm of random matrices. *The Annals of Probability 8*(2), 252–261.

Haff, L. R. (1991). The variational form of certain Bayes estimators. *The Annals of Statistics 19*(3), 1163–1190.

Huang, J. Z., N. Liu, M. Pourahmadi, and L. Liu (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika 93*(1), 85.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Stanford, California, United States, pp. 361–379.

Lam, C. and J. Fan (2007). Sparsistency and rates of convergence in large covariance matrices estimation. *Arxiv preprint arXiv:0711.3933*.

Lauritzen, S. (1996). *Graphical models*. Oxford University Press, USA.

Ledoit, O. and M. Wolf (2004a). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management 30*(4), 110–119.

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

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

Luenberger, D. G. (1998). *Investment science*. Oxford University Press New York.

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

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

Michaud, R. O. (1989). The Markowitz Optimization Enigma: Is Optimized Optimal. *Financial Analysts Journal 45*(1), 31–42.

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

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

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

Silverstein, J. W. (1985). The smallest eigenvalue of a large dimensional wishart matrix. *The Annals of Probability 13*(4), 1364–1368.

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

Stein, C. (1975). Estimation of a covariance matrix. *Reitz Lecture, IMS-ASA Annual Meeting (Also unpublished lecture notes)*.

Stein, C. (1977). Lectures on the theory of estimation of many parameters (In Russian). In I. Ibraguniv and M. Nikulin (Eds.), *Studies in the Statistical Theory of Estimation, Part I*, Proceedings of Scientific Seminars of the Steklov Institute, pp. 4–65. Leningrad Division 74.

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

Warton, D. I. (2008). Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices. *Journal of the American Statistical Association 103*(481), 340–349.

Won, J. H. and S.-J. Kim (2006). Maximum Likelihood Covariance Estimation with a Condition Number Constraint. In *Proceedings of the Fortieth Asilomar Conference on Signals, Systems and Computers*, pp. 1445–1449.

Wu, W. B. and M. Pourahmadi (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika 90*(4), 831.

Yang, R. and J. O. Berger (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics 22*(3), 1195–1211.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika 94*(1), 19–35.

A Uniqueness of the solution to (11)

The function $J^{(i)}_{\kappa_{\max}}(u)$ is convex and is constant on the interval $[1/(\kappa_{\max} l_i), 1/l_i]$, where l_i is the ith largest sample eigenvalue. It is strictly decreasing or increasing if $u < 1/(\kappa_{\max} l_i)$ or $u > 1/l_i$, respectively. Thus, the function $J_{\kappa_{\max}}(u) = \sum_{i=1}^{p} J^{(i)}_{\kappa_{\max}}(u)$ has a region on which it is a constant if and only if

$$\Big[1/(\kappa_{\max} l_1), 1/l_1\Big] \bigcap \Big[1/(\kappa_{\max} l_p), 1/l_p\Big] \neq \emptyset,$$

or equivalently, $1/\big(k_{\max} l_p\big) < 1/l_1$, *i.e.*, $\kappa_{\max} > \mathrm{cond}(S)$. This is precisely the condition so that the estimator reduces to the sample covariance matrix S. Therefore, provided that $\kappa_{\max} \leq \mathrm{cond}(S)$, the convex function $J_{\kappa_{\max}}(u)$ does not have a constant region, hence has the unique minimizer $u^\star$. On the other hand, if $\kappa_{\max} > \mathrm{cond}(S)$, the maximizer $u^\star$ is not unique but $\mu_i(u^\star) = l_i$ for every $i = 1, \ldots, p$. Hence, for this case, $\widehat{\Sigma}_{\mathrm{cond}} = S$ for all the maximizers.

B An algorithm for solving (11)

Without loss of generality, we assume that $\kappa_{\max} < l_1/l_p = \mathrm{cond}(S)$. As discussed in Appendix A, the function $J_{\kappa_{\max}}(u) = \sum_{i=1}^{p} J_{\kappa_{\max}}^{(i)}(u)$ is strictly decreasing for $u < 1/l_1$ and strictly increasing for $u \geq 1/(\kappa_{\max} l_p)$. Therefore, it suffices to consider $u \in \mathcal{I} = [1/l_1, 1/(\kappa_{\max} l_p))$.

Suppose an oracle tells us the values of α and β, the largest index such that $1/l_\alpha < u^\star$ and the smallest index such that $1/l_\beta > \kappa_{\max} u^\star$, respectively. Then,

$$J_{\kappa_{\max}}(u) = \sum_{i=1}^{\alpha} (l_i(\kappa_{\max} u) - \log(\kappa_{\max} u)) + \sum_{i=\alpha+1}^{\beta-1} (1 + \log l_i) + \sum_{i=\beta}^{p} (l_i u - \log u),$$

and the minimizer is immediately given by (12):

$$u^\star = \frac{\alpha + p - \beta + 1}{\sum_{i=1}^{\alpha} l_i + \sum_{i=\beta}^{p} \kappa_{\max} l_i}.$$

Now the problem is how to determine α and β. The main idea is that, for a fixed α and β, the value

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

coincides with $u^\star$ if and only if

$$1/l_\alpha < u_{\alpha,\beta} \leq 1/l_{\alpha+1} \tag{18}$$

and

$$1/l_{\beta-1} \leq \kappa_{\max} u_{\alpha,\beta} < 1/l_\beta. \tag{19}$$

The intersection of these two intervals is either empty or depending on the configuration of $l_1, \ldots, l_p, \kappa_{\max}$, one of the four intervals: $(1/l_\alpha, 1/(\kappa_{\max} l_\beta)]$, $(1/l_\alpha,\ 1/l_{\alpha+1}]$, $[1/(\kappa_{\max} l_{\beta-1}),$ $1/(\kappa_{\max} l_\beta))$, and $[1/(\kappa_{\max} l_{\beta-1}), 1/l_{\alpha+1}]$, the interior of which no other $1/l_i$ or $1/(\kappa_{\max} l_j)$ lies in. Starting from $1/l_1$, and by separately advancing the indexes for $1/l_i$ and $1/(\kappa_{\max} l_j)$, we can find α and β satisfying conditions (18) and (19), hence $u^\star$, in $O(p)$ operations. Algorithm 1 describes the procedure.

C 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

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

Algorithm 1 Solution method to the optimization problem (11)

Require: $l_1 \geq \ldots \geq l_p$, $1 < \kappa_{\max} < l_1/l_p$

$\alpha \leftarrow 1$, $\beta \leftarrow 2$, *lowerbound* $\leftarrow 1/l_1$
*has_lowerbound_factor_*$\kappa_{\max} \leftarrow$ true
loop
  if (*has_lowerbound_factor_*$\kappa_{\max}$) **then**
    while ($\beta \leq p$ and $1/(\kappa_{\max} l_\beta) \leq 1/l_\alpha$) **do** {*increase* β *until* $1/(\kappa_{\max} l_\beta) > 1/l_\alpha$}
      $\beta \leftarrow \beta + 1$
    end while
    if ($1/(\kappa_{\max} l_\beta) < 1/l_{\alpha+1}$) **then** {*case 1*}
      upperbound $\leftarrow 1/(\kappa_{\max} l_\beta)$
      *has_lowerbound_factor_*$\kappa_{\max} \leftarrow$ false
    else {*case 2*}
      upperbound $\leftarrow 1/l_{\alpha+1}$
      $\alpha \leftarrow \alpha + 1$
      *has_lowerbound_factor_*$\kappa_{\max} \leftarrow$ true
    end if
  else
    while ($1/l_{\alpha+1} \leq 1/(\kappa_{\max} l_{\beta-1})$) **do** {*increase* α *until* $1/l_{\alpha+1} > 1/(\kappa_{\max} l_{\beta-1})$}
      $\alpha \leftarrow \alpha + 1$
    end while
    if ($1/(\kappa_{\max} l_\beta) < 1/l_{\alpha+1}$) **then** {*case 3*}
      upperbound $\leftarrow 1/(\kappa_{\max} l_\beta)$
      $\alpha \leftarrow \alpha + 1$
      *has_lowerbound_factor_*$\kappa_{\max} \leftarrow$ false
    else {*case 4*}
      upperbound $\leftarrow 1/l_{\alpha+1}$
      *has_lowerbound_factor_*$\kappa_{\max} \leftarrow$ true
    end if
  end if
  $u_{\alpha,\beta} \leftarrow (\alpha + p - \beta + 1)/(\sum_{i=1}^{\alpha} l_i + \sum_{i=\beta}^{p} \kappa_{\max} l_i)$
  if (*lowerbound* $\leq u^\star_{\alpha,\beta} \leq$ *upperbound*) **then**
    $u^\star \leftarrow u_{\alpha,\beta}$
  end if
  lowerbound $\leftarrow$ *upperbound* {*proceed to the next interval*}
end loop

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

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

and

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

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

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

   and

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

   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, $u^\star(\nu) \leq u^\star(\nu_0)$.
   Then, 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, $v^\star(\nu) \geq v^\star(\nu_0)$.
Then, 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.

D Proof of Theorem 1

Suppose the spectral decomposition of the k-th fold covariance matrix estimate $\widehat{\Sigma}_\nu^{[-k]}$, with $\kappa_{\max} = \nu$, is

$$\widehat{\Sigma}_\nu^{[-k]} = Q^{[-k]} \mathrm{diag}\big(\widehat{\lambda}_1^{[-k]}, \ldots, \widehat{\lambda}_p^{[-k]}\big) \big(Q^{[-k]}\big)^T$$

with

$$\widehat{\lambda}_i^{[-k]} = \begin{cases} v^{[-k]*} & \text{if } l_i^{[-k]} < v^{[-k]*} \\ l_i^{[-k]} & \text{if } v^{[-k]*} \leq l_i^{[-k]} < \nu v^{[-k]*} \\ \nu v^{[-k]*} & \text{if } l_i^{[-k]} \geq \nu v^{[-k]*}, \end{cases}$$

where $l_i^{[-k]}$ is the i-th largest eigenvalue of the k-th fold sample covariance matrix $S^{[-k]}$, and $v^{[-k]*}$ is obtained according to the method described in Section 2. Since $\widehat{\Sigma}_\nu^{[-k]} = S^{[-k]}$ if $\nu \geq l_1^{[-k]}/l_p^{[-k]} = \mathrm{cond}(S^{[-k]})$,

$$\widehat{\kappa}_{\max} \leq \max_{k=1,\ldots,K} l_1^{[-k]}/l_p^{[-k]}. \tag{20}$$

The right hand side of (20) converges in probability to the condition number κ of the true covariance matrix, as n increases while p is fixed. Hence,

$$\lim_{n\to\infty} P\Big(\widehat{\kappa}_{\max} \leq \kappa\Big) = 1.$$

We now prove that

$$\lim_{n\to\infty} P\Big(\widehat{\kappa}_{\max} \geq \kappa\Big) = 1.$$

by showing that $\widehat{\mathrm{PR}}\big(\nu\big)$ is an asymptotically decreasing function in ν.

Recall that

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

where

$$l_k\big(\widehat{\Sigma}_\nu^{[-k]}, X_k\big) \;=\; -(n_k/2)\Big[\,\mathbf{Tr}\left\{\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1} X_k X_k^T/n_k\right\} - \log\det\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1}\Big],$$

which, by the definition of $\widehat{\Sigma}_\nu^{[-k]}$, is everywhere differentiable but at a finite number of points.

To see the asymptotic monotonicity of $\widehat{\mathrm{PR}}(\nu)$, consider the derivative $-\partial l_k\big(\widehat{\Sigma}_\nu^{[-k]}, X_k\big)/\partial\nu$:

$$\begin{aligned}
-\frac{\partial l_k\big(\widehat{\Sigma}_\nu^{[-k]}, X_k\big)}{\partial\nu} &= \frac{n_k}{2}\left[\,\mathbf{Tr}\left(\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1}\frac{\partial\widehat{\Sigma}_\nu^{[-k]}}{\partial\nu}\right) + \mathbf{Tr}\left\{\frac{\partial\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1}}{\partial\nu}\Big(X_k X_k^T/n_k\Big)\right\}\right] \\
&= \frac{n_k}{2}\left[\,\mathbf{Tr}\left(\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1}\frac{\partial\widehat{\Sigma}_\nu^{[-k]}}{\partial\nu}\right) + \mathbf{Tr}\left\{\frac{\partial\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1}}{\partial\nu}\widehat{\Sigma}_\nu^{[-k]}\right\}\right. \\
&\quad \left. + \mathbf{Tr}\left\{\frac{\partial\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1}}{\partial\nu}\Big(X_k X_k^T/n_k - \widehat{\Sigma}_\nu^{[-k]}\Big)\right\}\right] \\
&= \frac{n_k}{2}\left[\frac{\partial}{\partial\nu}\,\mathbf{Tr}\left(\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1}\widehat{\Sigma}_\nu^{[-k]}\right) + \mathbf{Tr}\left\{\frac{\partial\big(\widehat{\Sigma}_\nu^{[-k]}\big)^{-1}}{\partial\nu}\Big(X_k X_k^T/n_k - \widehat{\Sigma}_\nu^{[-k]}\Big)\right\}\right] \\
&= \frac{n_k}{2}\,\mathbf{Tr}\left\{\frac{\partial\tilde{\Sigma}_\nu^{-1}}{\partial\nu}\Big(X_k X_k^T/n_k - \widehat{\Sigma}_\nu^{[-k]}\Big)\right\}.
\end{aligned}$$

As n and n_k increases, $\widehat{\Sigma}_\nu^{[-k]}$ converges almost surely to the inverse of the solution to the following optimization problem

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(\Omega\Sigma) - \log\det\Omega \\ \text{subject to} & \mathrm{cond}(\Omega) \le \nu, \end{array}$$

with Σ and ν replacing S and $\kappa_{\max}$ in (7). We denote the limit of $\widehat{\Sigma}_\nu^{[-k]}$ by $\tilde{\Sigma}_\nu$. For the spectral decomposition of Σ

$$\Sigma = R\ \mathrm{diag}\big(\lambda_1, \ldots, \lambda_p\big) R^T, \tag{21}$$

$\tilde{\Sigma}_\nu$ is given as

$$\tilde{\Sigma}_\nu = R\ \mathrm{diag}\big(\psi_1(\nu), \ldots, \psi_p(\nu)\big)\ R^T, \tag{22}$$

where, for some $\tau(\nu) > 0$,

$$\psi_i(\nu) = \begin{cases} \tau(\nu) & \text{if} \quad \lambda_i \le \tau(\nu) \\ \lambda_i & \text{if} \quad \tau(\nu) < \lambda_i \le \nu\tau(\nu) \\ \nu\tau(\nu), & \text{if} \quad \nu\tau(\nu) < \lambda_i. \end{cases}$$

Recall from Proposition 1 that $\tau(\nu)$ is decreasing in ν and $\nu\tau(\nu)$ is increasing.

Let c_k be the limit of $n_k/(2n)$ when both n and n_k increases. Then, $X_k X_k^T/n_k$ converges almost surely to Σ. Thus,

$$-\frac{1}{n}\frac{\partial l_k\big(\widehat{\Sigma}_\nu^{[-k]}, X_k\big)}{\partial \nu} \to c_k \,\mathbf{Tr}\left\{\frac{\partial \tilde{\Sigma}_\nu^{-1}}{\partial \nu}\Big(\Sigma - \tilde{\Sigma}_\nu\Big)\right\}, \quad \text{almost surely.} \tag{23}$$

We now study (23). First, if $\nu \geq \kappa$, then $\tilde{\Sigma}_\nu = \Sigma$, and the RHS of (23) degenerates to 0. Now we consider the non-trivial case that $\nu < \kappa$. From (22),

$$\frac{\partial \tilde{\Sigma}_\nu^{-1}}{\partial \nu} = R\frac{\partial \Psi^{-1}}{\partial \nu}R^T = R\ \mathrm{diag}\left(\frac{\partial \psi_1^{-1}}{\partial \nu}, \ldots, \frac{\partial \psi_p^{-1}}{\partial \nu}\right)R^T,$$

where

$$\frac{\partial \psi_i^{-1}}{\partial \nu} = \begin{cases} -\frac{1}{\tau(\nu)^2}\frac{\partial \tau(\nu)}{\partial \nu} & (\geq 0) \quad \text{if} \quad \lambda_i \leq \tau(\nu) \\ 0 & \quad\quad\ \ \text{if} \quad \tau(\nu) < \lambda_i \leq \nu\tau(\nu) \\ -\frac{1}{\nu^2\tau(\nu)^2}\frac{\partial(\nu\tau(\nu))}{\partial \nu} & (\leq 0) \quad \text{if} \quad \nu\tau(\nu) < \lambda_i. \end{cases}$$

From (21) and (22),

$$\Sigma - \tilde{\Sigma}_\nu = R\mathrm{diag}\big(\lambda_1 - \psi_1, \ldots, \lambda_p - \psi_p\big)R^T,$$

where

$$\lambda_i - \psi_i = \begin{cases} \lambda_i - u(\nu) & (\leq 0) \quad \text{if} \quad \lambda_i \leq \tau(\nu) \\ 0 & \quad\quad\ \ \text{if} \quad \lambda_i \leq \nu\tau(\nu) \\ \lambda_i - \nu u(\nu) & (\geq 0) \quad \text{if} \quad \nu\tau(\nu) < \lambda_i. \end{cases}$$

Thus, the RHS of (23) is less than 0 and the almost sure limit of $\widehat{\mathrm{PR}}(\nu)$ is decreasing in ν.

Finally, from the monotonicity $\widehat{\mathrm{PR}}(\widehat{\kappa}_{\max}) \leq \widehat{\mathrm{PR}}(\kappa)$, we conclude that

$$\lim_{n\to\infty} P\left(\widehat{\kappa}_{\max} \geq \kappa\right) = 1.$$

E Proof of Theorem 2

(i) Suppose the spectral decomposition of S is QLQ^T, with Q orthogonal and $L = \mathrm{diag}(l_1, \ldots, l_p)$, as given in (1). Then the solution to (7) for the given u is represented as $\widehat{\Sigma}(u) = Q\widehat{\Lambda}^{-1}Q^T$, where $\widehat{\Lambda}^{-1} = \mathrm{diag}(\widehat{\lambda}_1^{-1}, \ldots, \widehat{\lambda}_p^{-1})$, with

$$\widehat{\lambda}_i^{-1} = \begin{cases} 1/(\kappa_{\max}u), & \text{if } \ l_i \leq 1/(\kappa_{\max}u) \\ l_i, & \text{if } 1/(\kappa_{\max}u) \leq l_i < 1/u \\ 1/u, & \text{if } \ l_i \geq 1/u. \end{cases}$$

for $i = 1, \ldots, p$. The conditional risk of $\widehat{\Sigma}(u)$, given the sample eigenvalues $\underline{l} = (l_1, \ldots, l_p)$, is

$$\mathbf{E}\left(\mathcal{L}_{\mathrm{ent}}(\widehat{\Sigma}(u), \Sigma)\big|\underline{l}\right) \ = \ \sum_{i=1}^{p}\left\{\widehat{\lambda}_i^{-1}\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log\widehat{\lambda}_i^{-1}\right\} + \log\det\Sigma - p,$$

where $a_{ii}(Q) = \sum_{j=1}^{p} q_{ji}^2 \lambda_j^{-1}$ and q_{ji} is the (j,i)-th element of the orthogonal matrix Q. This is because

$$\begin{aligned}\mathcal{L}_{\text{ent}}\big(\widehat{\Sigma}(u), \Sigma\big) &= \mathbf{Tr}\left(\widehat{\Lambda}^{-1} A(Q)\right) - \log\det\left(\widehat{\Lambda}^{-1}\right) + \log\det\Sigma - p \\ &= \sum_{i=1}^{p} \left\{\widehat{\lambda}_i^{-1} a_{ii}(Q) - \log\widehat{\lambda}_i^{-1}\right\} + \log\det\Sigma - p, \end{aligned} \tag{24}$$

where $A(Q) = Q^T \Sigma^{-1} Q$.

In (24), the summand has the form

$$x\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log x$$

whose minimum is achieved at $x = 1/\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right)$. Since $\sum_{j=1}^{p} q_{ji}^2 = 1$, and $\Sigma^{-1} \in \mathcal{D}\big(\kappa_{\max}, u\big)$ if $\Sigma \in \mathcal{D}\big(\kappa_{\max}, u\big)$, we have $u \le a_{ii}(Q) \le \kappa_{\max} u$. Hence $1/\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right)$ lies between $1/u$ and $1/\kappa_{\max} u$ almost surely. Therefore,

1. If $l_i \le 1/(\kappa_{\max} u)$, then $\widehat{\lambda}_i^{-1} = 1/(\kappa_{\max} u)$ and

$$\widehat{\lambda}_i^{-1}\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log\widehat{\lambda}_i^{-1} \le l_i\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log l_i.$$

2. If $1/\kappa_{\max} u \le l_i < 1/u$, then $\widehat{\lambda}_i^{-1} = l_i$ and

$$\widehat{\lambda}_i^{-1}\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log\widehat{\lambda}_i^{-1} = l_i\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log l_i.$$

3. If $l_i \ge 1/u$,then $\widehat{\lambda}_i = 1/u$ and

$$\widehat{\lambda}_i^{-1}\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log\widehat{\lambda}_i^{-1} \le l_i\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log l_i.$$

Thus,

$$\sum_{i=1}^{p} \left\{\widehat{\lambda}_i^{-1}\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log\widehat{\lambda}_i^{-1}\right\} \le \sum_{i=1}^{p} \left\{l_i\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log l_i\right\}$$

and the risk with respect to the entropy loss is

$$\begin{aligned}\mathcal{R}_{\text{ent}}\big(\widehat{\Sigma}(u)\big) &= \mathbf{E}\left[\sum_{i=1}^{p} \left\{\widehat{\lambda}_i^{-1}\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log\widehat{\lambda}_i^{-1}\right\}\right] \\ &\le \mathbf{E}\left[\sum_{i=1}^{p} \left\{l_i\,\mathbf{E}\left(a_{ii}(Q)\big|\underline{l}\right) - \log l_i\right\}\right] \\ &= \mathcal{R}_{\text{ent}}\big(S\big). \end{aligned}$$

In other words, $\widehat{\Sigma}(u)$ has a smaller risk than S, provided $\underline{\lambda}^{-1} \in \mathcal{D}(\kappa_{\max}, u)$.

(ii) Suppose the true covariance matrix Σ has the spectral decomposition $\Sigma = R\Lambda R^T$ with R orthogonal and $\Lambda = \mathrm{diag}\big(\lambda_1, \ldots, \lambda_p\big)$. Let $A = R\Lambda^{1/2}$, then S has the same distribution as AS_0A^T, where S_0 is the sample covariance matrix when the true covariance matrix is the identity. Recall the operational definition of the largest eigenvalue l_1 of S

$$l_1 = \max_{v \neq 0} \frac{v^T S v}{v^T v} = \max_{w \neq 0} \frac{w^T S_0 w}{w^T \Lambda^{-1} w},$$

where $w = A^T v$. In addition to this, since for any $w \neq 0$,

$$\lambda_1^{-1} = \min_{w \neq 0} \frac{w^T \Lambda^{-1} w}{w^T w} \leq \frac{w^T \Lambda^{-1} w}{w^T w},$$

we have

$$l_1 \leq \lambda_1 \max_{w \neq 0} \frac{w^T S_0 w}{w^T w} = \lambda_1 e_1, \tag{25}$$

where e_1 is the largest eigenvalue of S_0. Using essentially the same argument, we can show that

$$l_p \geq \lambda_p e_p, \tag{26}$$

where e_p is the smallest eigenvalue of S_0. Then, from the results by Geman (1980) and Silverstein (1985), we see that

$$P\left(\left\{e_1 \leq \big(1 + \sqrt{\gamma}\big)^2,\ e_p \geq \big(1 - \sqrt{\gamma}\big)^2\right\} \text{ eventually}\right) = 1. \tag{27}$$

The combination of (25)–(27) leads to

$$P\left(\left\{l_1 \leq \lambda_1\big(1 + \sqrt{\gamma}\big)^2,\ l_p \geq \lambda_p\big(1 - \sqrt{\gamma}\big)^2\right\} \text{ eventually}\right) = 1.$$

On the other hand, if $\kappa_{\max} \geq \kappa\big(1 - \sqrt{\gamma}\big)^{-2}$, then

$$\left\{l_1 \leq \lambda_1\big(1 + \sqrt{\gamma}\big)^2,\ l_p \geq \lambda_p\big(1 - \sqrt{\gamma}\big)^2\right\} \subset \left\{\max\left(\frac{l_1}{\lambda_p}, \frac{\lambda_1}{l_p}\right) \leq \kappa_{\max}\right\}.$$

Also, if $\max\big(l_1/\lambda_p,\ \lambda_1/l_p\big) \leq \kappa_{\max}$, then

$$1/(\kappa_{\max}\lambda_p) \leq 1/l_1 \leq 1/(\kappa_{\max} l_p) \leq 1/\lambda_1.$$

Since, from Appendix B, $u^\star$ lies between $(1/l_1)$ and $1/(\kappa_{\max} l_p)$,

$$u^\star \leq \lambda_1^{-1} \quad \text{and} \quad \lambda_p^{-1} \leq \kappa_{\max} u^\star.$$

Therefore,

$$\left\{\max\left(\frac{l_1}{\lambda_p}, \frac{\lambda_1}{l_p}\right) \leq \kappa_{\max}\right\} \subset \left\{\Sigma \in \mathcal{D}\big(\kappa_{\max}, u^\star\big)\right\},$$

which concludes the proof.

F Proof of Proposition 2

We are given that

$$\pi(\lambda_1, \lambda_1, \ldots, \lambda_p) = \mathrm{e}^{-g_{\max}\frac{\lambda_1}{\lambda_p}} \qquad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0.$$

Now

$$\int_C \pi(\lambda_1, \lambda_1, \ldots, \lambda_p) d\underset{\sim}{\lambda} = \int_C \mathrm{e}^{-g_{\max}\frac{\lambda_1}{\lambda_p}} d\underset{\sim}{\lambda},$$

where $C = \{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0\}$.

Let us now make the following change of variables: $x_i = \lambda_i - \lambda_{i+1}$ for $i = 1, 2, .., p-1$, and $x_p = \lambda_p$. The inverse transformation yields $\lambda_i = \sum_{j=i}^{p} x_j$ for $i = 1, 2, .., p$. It is straightforward to verify that the Jacobian of this transformation is given by $|J| = 1$.

Now we can therefore rewrite the integral above as

$$\begin{aligned}
\int_C \mathrm{e}^{-g_{\max}\left(\frac{\lambda_1}{\lambda_p}\right)} d\underset{\sim}{\lambda} &= \int_{\mathbb{R}_1^p} \mathrm{e}^{-g_{\max}\left(\frac{x_1+x_2+..+x_p}{x_p}\right)} dx_1 dx_2 \cdots dx_p \\
&= \mathrm{e}^{-g_{\max}} \int \left[\prod_{i=1}^{p-1} \int \mathrm{e}^{-g_{\max}\left(\frac{x_i}{x_p}\right)} dx_i \right] dx_p \\
&= \mathrm{e}^{-g_{\max}} \int_0^\infty \left(\frac{x_p}{g_{\max}}\right)^{p-1} dx_p \\
&= \frac{\mathrm{e}^{-g_{\max}}}{g_{\max}^{p-1}} \int_0^\infty x_p^{p-1} dx_p \\
&= \infty.
\end{aligned}$$

To prove that the posterior yields a proper distribution we proceed as follows:

$$\begin{aligned}
&\int_C \pi(\underset{\sim}{\lambda}) f(\underset{\sim}{\lambda}, \underset{\sim}{l}) d\underset{\sim}{\lambda} \\
&\propto \int_C \exp\left(-\frac{n}{2}\sum_{i=1}^{p}\frac{l_i}{\lambda_i}\right)\left(\prod_{i=1}^{p}\lambda_i\right)^{-\frac{n}{2}} \mathrm{e}^{-g_{\max}\left(\frac{\lambda_1}{\lambda_p}\right)} d\underset{\sim}{\lambda} \\
&\leq \int_C \exp\left(-\frac{n}{2}\sum_{i=1}^{p}\frac{l_p}{\lambda_i}\right)\left(\prod_{i=1}^{p}\lambda_i\right)^{-\frac{n}{2}} \mathrm{e}^{-g_{\max}\left(\frac{\lambda_1}{\lambda_p}\right)} d\underset{\sim}{\lambda} \;\; as \;\; l_p \leq l_i \;\; \forall i = 1, 2, \ldots, p \\
&\leq \int_C \exp\left(-\frac{n}{2}\sum_{i=1}^{p}\frac{l_p}{\lambda_i}\right)\left(\prod_{i=1}^{p}\lambda_i\right)^{-\frac{n}{2}} \mathrm{e}^{-g_{\max}} d\underset{\sim}{\lambda} \;\; as \;\; \frac{\lambda_1}{\lambda_p} \geq 1 \\
&\leq \mathrm{e}^{-g_{\max}} \prod_{i=1}^{p}\left(\int_0^\infty \mathrm{e}^{-\frac{n}{2}\frac{l_p}{\lambda_i}} \lambda_i^{-\frac{n}{2}}\right).
\end{aligned}$$

The above integrand is the density of the inverse Gamma distribution and therefore the corresponding integral above has a finite normalizing constant and thus yielding a proper posterior.