Regularizing Sample Estimates of Covariance Matrices by Condition Number

Joong-Ho Won

Division of Biostatistics

Stanford University, Stanford, CA 94305

Johan Lim

Department of Statistics

Seoul National University, Seoul, Korea

Seung-Jean Kim

Citi Capital Advisors, New York City, NY

Bala Rajaratnam

Department of Statistics

Stanford University, Stanford, CA 94305

July 4, 2010

Author's Footnote:

Joong-Ho Won is Postdoctoral Scholar, Division of Biostatistics, Stanford University, Stanford, CA 94305 (e-mail: jhwon@stanford.edu); Johan Lim is Associate Professor, Department of Statistics, Seoul National University, Seoul, Korea. (e-mail: johanlim@snu.ac.kr; Seung-Jean Kim is Vice President, Citi Capital Advisors, New York City, NY (e-mail: seungjean@gmail.com); and Bala Rajaratnam is Assistant Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (e-mail: brajarat@stanford.edu). JHW was partially supported by NSF grant CCR 0309701 and NIH MERIT Award R37EB02784. BR was supported in part by NSF grant DMS 0505303. The authors thank Professors Charles Stein, Richard A. Olshen, Robert M. Gray and Dr. Alessandro Magnani for helpful comments and suggestions.

Abstract

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

Keywords: covariance estimation, regularization, convex optimization, eigenvalue, cross-validation, two-sample testing, portfolio optimization

1 Introduction

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

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

maximizes the log-likelihood of a zero-mean p-variate Gaussian distribution

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

where $|\Sigma|$ and $\det\Sigma$ both denote the determinant of Σ, $\mathbf{Tr}(A)$ denotes the trace of A (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, 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, for *e.g.*, two-sample testing and mean-variance portfolio theory. Even when $n > p$, the eigenstructure tends to be systematically distorted unless p/n is extremely small, resulting in numerically ill-conditioned estimators for Σ; see Dempster (1972) and Stein (1975). In multivariate two-sample testing, most test statistics involve inverting S, making them incalculable for $n < p$. In mean-variance portfolio optimization (Luenberger, 1998; Markowitz, 1952), an ill-conditioned covariance matrix may amplify estimation error present in the mean return estimate (Ledoit and Wolf, 2004a; Michaud, 1989). A common approach to mitigate the problem of numerical stability is regularization.

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

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

where $\text{cond}(M)$ stands for the condition number, a measure of numerical stability, of a matrix M. M is invertible if $\text{cond}(M)$ is finite, and is ill-conditioned if $\text{cond}(M)$ is finite but high. By bounding the condition number of the sample estimate by a regularization parameter $\kappa_{\max}$, we address the problem of invertibility or ill-conditioning directly. It turns out that the resulting regularized matrix falls into a broad family of Steinian type shrinkage estimators, which shrink the eigenvalues of the sample covariance matrix towards a given structure (James and Stein, 1961; Stein, 1956).

Numerous authors have explored alternative estimators for Σ (or Σ^{-1}) that perform better than the sample covariance estimator S from a decision-theoretic point of view. Many of these estimators give substantial risk reductions compared to S in small sample sizes. Most often these estimators are Steinian shrinkage estimators. A simple example is the 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 norm risk. 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). To list a few nonlinear Steinian estimators, James and Stein (1961) study a constant risk minimax estimator and

its modification in a class of orthogonally invariant estimators. Dey and Srinivasan (1985) provide another minimax estimator which dominates the James-Stein estimator. Yang and Berger (1994) and Daniels and Kass (2001) consider a reference prior and hierarchical priors respectively, that yield posterior shrinkage toward a specified structure.

Likelihood-based approaches using multivariate Gaussian models have provided different perspectives to the regularization problem. Warton (2008) derives a novel family of linear shrinkage estimators from a penalized maximum likelihood framework. This formulation enables cross-validating the regularization parameter. He also studies the application of his estimator to testing equality of means of two populations. Related work in the area include Sheena and Gupta (2003). An extensive literature review is not undertaken here but we note that the approaches mentioned above fall in the class of covariance estimation problems which do not assume or impose sparsity, on either the covariance matrix or its inverse.

1.1 Regularization by shrinking sample eigenvalues

We briefly review Steinian 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

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

where $\text{diag}(l_1, \ldots, l_p)$ 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 regularizes S by transforming its eigenvalues:

$$\widehat{\Sigma} = Q\text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)Q^T. \tag{4}$$

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

Many regularization schemes for covariance matrices rely explicitly or implicitly on the concept of shrinking the eigenvalues of the sample covariance matrix. In the linear shrinkage estimator

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

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

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

(If F does not commute with S, it does not have the form (4).) In Stein's estimator (Stein, 1975, 1977, 1986), $\widehat{\lambda}_i$ are obtained by applying isotonic regression (Lin and Perlman, 1985) to l_i/d_i, $i = 1, \ldots, p$ with $d_i = \big(n - p + 1 + 2l_i \sum_{j \neq i}(l_i - l_j)^{-1}\big)/n$, in order to maintain the nonincreasing order constraint. In the penalized 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 ($\widehat{\lambda}_i = \min\{l_i, U\}$). By focusing on only one of the two ends of the eigenspectrum, the 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 (L or U) needs to be investigated.

1.2 Regularization by imposing a condition number constraint

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

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

where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are the maximum and the minimum eigenvalues of Σ, respectively.

The likelihood approach to regularizing the sample estimate of a covariance matrix by constraining its condition number (2) can then be formulated as

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

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

The covariance estimation problem (6) 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 regularized covariance matrix $\widehat{\Sigma}_{\text{cond}}$ that solves (6) has a Steinian shrinkage form in (4) as

$$\widehat{\Sigma}_{\text{cond}} = Q\text{diag}(\widehat{\lambda}_i, \ldots, \widehat{\lambda}_p)Q^T, \tag{7}$$

with the eigenvalues

$$\widehat{\lambda}_i = \min\big(\max(\tau^\star, l_i), \kappa_{\max}\tau^\star\big), \tag{8}$$

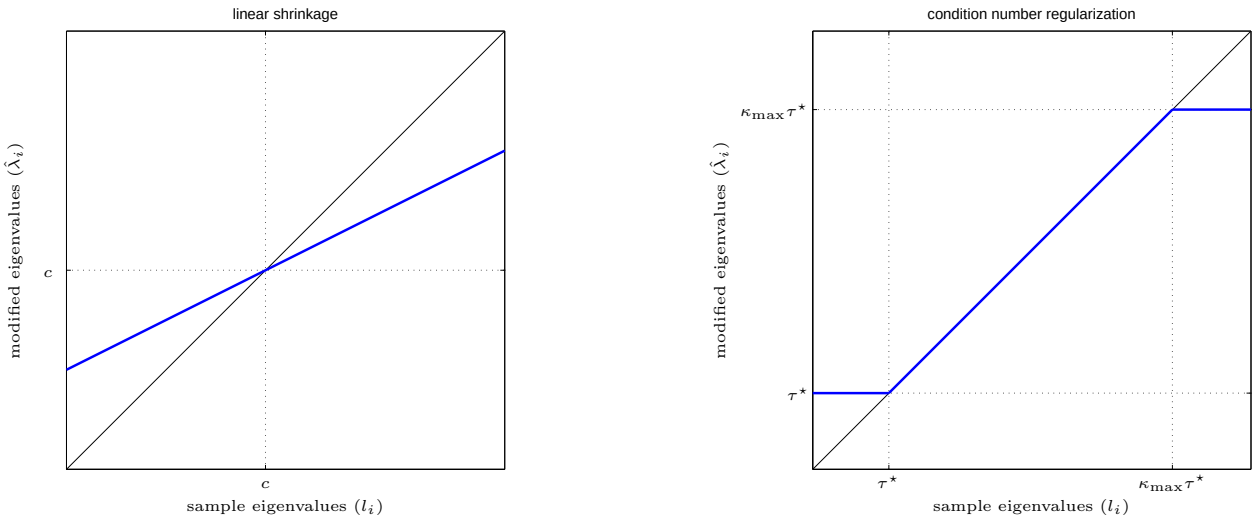


Figure 1: Comparison of eigenvalue shrinkage of the linear shrinkage estimator (left) and the condition number-constrained estimator (right).

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 (8) scales well to much larger size estimation problems, compared with standard solution methods for (6).

The nonlinear transform (8) 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 $\tau^\star$, where $\tau^\star$ is determined by the data and the choice of the regularization parameter $\kappa_{\max}$. Figure 1 illustrates the transform (8) in comparison with that of the linear shrinkage estimator.

An important issue in this regularization scheme is the selection of $\kappa_{\max}$. We propose a selection procedure that minimizes predictive risk approximated using cross-validation. 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 increases. The variance of $\widehat{\kappa}_{\max}$ decreases when either n or

p increases.

1.3 The outline

In the next section, we derive the regularized covariance matrix as a solution to the maximum likelihood estimation problem (6). We offer a solution algorithm and provide a geometric interpretation of the derived estimator, elucidating properties of the estimator. We also show that the regularized covariance matrix dominates asymptotically the sample covariance matrix for a chosen risk function. 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. We illustrate two applications of the proposed regularization scheme. In Section 5, we explore its use for inference in two-sample problems. In Section 6, we describe the application in mean-variance portfolio optimization. Finally, we give our conclusions in Section 7. The proofs of the theoretical results discussed in the text are collected in the appendices.

2 Estimation of the condition number-regularized covariance matrices

This section gives the details of the solution (7) and shows how to compute $\tau^\star$ given $\kappa_{\max}$. It suffices to consider the case $\kappa_{\max} < l_1/l_p = \text{cond}(S)$, since otherwise the solution to (6) reduces to the sample covariance matrix S.

2.1 Derivation

It is well known that the log-likelihood (1) 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 (6) is equivalent to

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(\Omega S) - \log\det\Omega \\ \text{subject to} & uI \preceq \Omega \preceq \kappa_{\max} uI, \end{array} \tag{9}$$

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 the sample covariance matrix $S = QLQ^T$, with $L = \mathrm{diag}(l_1, \ldots, l_p)$ and $l_1 \geq \ldots \geq l_p \geq 0$. 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 (9) 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 (9)

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{p}(l_i\mu_i - \log\mu_i) \\ \text{subject to} & u \leq \mu_i \leq \kappa_{\max} u, \quad i = 1, \ldots, p, \end{array} \tag{10}$$

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 (10). The solution to (9) is then

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

We can reduce (10) to an equivalent *univariate* convex problem. We start by observing that (10) is equivalent to

$$\text{minimize} \quad \textstyle\sum_{i=1}^p \min_{u \le \mu_i \le \kappa_{\max} u}(l_i\mu_i - \log \mu_i). \tag{11}$$

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 (11) is given as

$$\mu_i^\star(u) = \operatorname*{argmin}_{u \le \mu_i \le \kappa_{\max} u} (l_i\mu_i - \log \mu_i) = \min\big\{ \max\{u, 1/l_i\}, \kappa_{\max} u\big\}. \tag{12}$$

Then (10) reduces to an unconstrained, univariate optimization problem

$$\text{minimize} \ \sum_{i=1}^p J^{(i)}_{\kappa_{\max}}(u), \tag{13}$$

where

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

This problem is convex, since each $J^{(i)}_{\kappa_{\max}}$ is convex in u. It follows that we can express the solution explicitly:

Theorem 1. *Provided that* $\kappa_{\max} < \mathrm{cond}(S)$, (13) *has the unique solution*

$$u^\star = \frac{\alpha + p - \beta + 1}{\sum_{i=1}^{\alpha} l_i + \sum_{i=\beta}^{p} \kappa_{\max} l_i}, \tag{14}$$

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* $1/l_\beta > \kappa_{\max} u^\star$. α *and* β *are not determined* a priori *but can be found in* $O(p)$ *operations on the sample eigenvalues* $l_1 \geq \ldots \geq l_p$. *If* $\kappa_{\max} > \mathrm{cond}(S)$, *the maximizer* $u^\star$ *is not unique but* $\widehat{\Sigma}_{\mathrm{cond}} = S$ *for all the maximizers.*

Proof. Given in Supplemental Section A. □

From the solution $u^\star$ to (13), we can write the solution (7) as

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

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 (6). The eigenvalues of this solution have the form (8), 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 and) truncated eigenvalues, in which 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 estimation of the condition number-regularized covariance matrices 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 method 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 (9) provides an intuitive geometric perspective to the original problem. Consider a function

$$J(u,v) = \min_{uI \preceq \Omega \preceq vI} (\mathbf{Tr}(\Omega S) - \log\det\Omega) \tag{15}$$

defined as the minimum of the objective of (9) over a fixed range $uI \preceq \Omega \preceq vI$, where $0 < u \leq v$. Following the argument that leads to (11), 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

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

Comparing this to (12), we observe that $\Omega^\star$, which achieves the minimum in (15), 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 (9): the solution $u^\star$ to (9) 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.

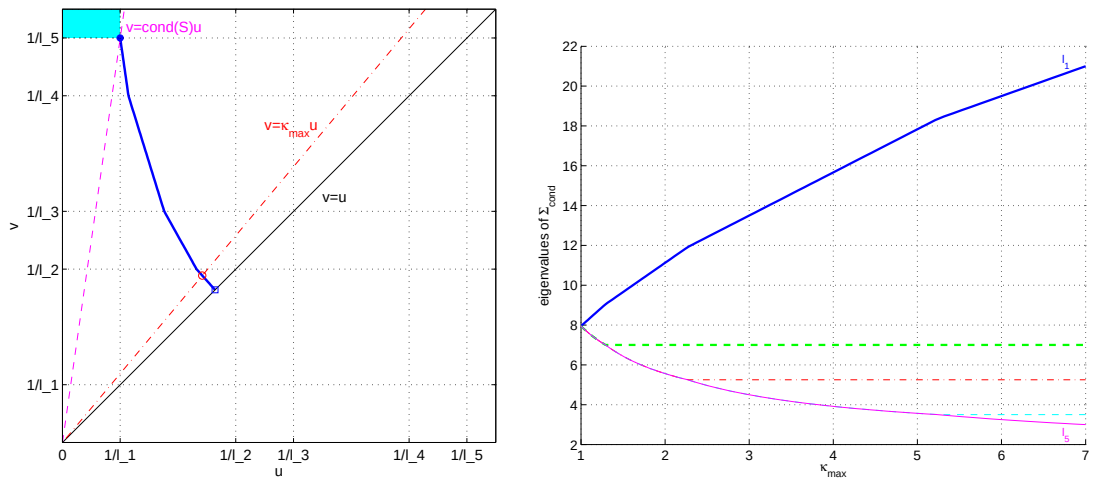
Proposition 1. *$u^\star(\kappa_{\max})$ is nonincreasing in $\kappa_{\max}$ and $v^\star(\kappa_{\max}) \triangleq \kappa_{\max} u^\star(\kappa_{\max})$ is nondecreasing, both almost surely.*

Proof. Given in Supplemental Section B. □

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, for $\tilde{\kappa}_{\max} > \kappa_{\max}$, the new optimal point $(u^\star(\tilde{\kappa}_{\max}), v^\star(\tilde{\kappa}_{\max}))$ lies on the line segment between the two points expressed in terms of the previous optimal point $(u^\star(\kappa_{\max}), v^\star(\kappa_{\max}))$:

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

The proposition also implies 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



(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 $\kappa_{\max}$.

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 eigenvalues are $(21, 7, 5.25, 3.5, 3)$. Here a point on the path represents the minimizer of $J(u, v)$ on a line $v = \kappa_{\max} u$ (hollow circle). The path starts from a point on the solid line $v = u$ ($\kappa_{\max} = 1$, square) and ends at $(1/l_1, 1/l_p)$, where the dashed line $v = \mathrm{cond}(S)u$ passes ($\kappa_{\max} = \mathrm{cond}(S)$, solid circle). Note that the starting point corresponds to $\widehat{\Sigma}_{\mathrm{cond}} = \gamma I$ for some $\gamma > 0$ 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 we see that the truncation ranges of the eigenvalues are nested.

2.3 Risk property

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

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

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 $\Lambda = \text{diag}\big(\lambda_1, \ldots, \lambda_p\big)$. 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 $\kappa = \lambda_1/\lambda_p$.

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

For $\gamma < 1$, if the true covariance matrix of the samples has a finite condition number, we can show that for a properly chosen $\kappa_{\max}$, $\widehat{\Sigma}_{\text{cond}} = \widehat{\Sigma}(u^\star)$ dominates the sample covariance matrix asymptotically.

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

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

Then, the following results hold.

(i) For a true covariance matrix $\Sigma \in \mathcal{D}(\kappa_{\max}, u)$, $\widehat{\Sigma}(u)$, which solves (9) for the given u (and $\kappa_{\max}$), has a smaller risk than the sample covariance matrix S with respect to the entropy loss.

(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 (13) for the given $\kappa_{\max}$.

Proof. Given in Supplemental Section C. □

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 an optimal $\kappa_{\max}$.

3.1 Predictive risk selection procedure

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

$$\mathrm{PR}(\nu) = \mathbf{E}\left[\mathbf{E}_{\tilde{X}}\Big\{\ \mathbf{Tr}(\widehat{\Sigma}_\nu^{-1}\tilde{X}\tilde{X}^T) - \log\det\widehat{\Sigma}_\nu^{-1}\Big\}\right], \tag{16}$$

where $\widehat{\Sigma}_\nu$ is the estimated condition number-regularized covariance matrix given independent samples $x_1, \cdots, x_n$ from a zero-mean Gaussian distribution on $\mathbb{R}^p$, with the value of the regularization parameter $\kappa_{\max}$ set to ν, and $\tilde{X} \in \mathbb{R}^p$ is a random vector, independent of the given samples, from the same distribution. We approximate the predictive risk using K-fold cross validation. K-fold cross validation divides the data matrix $\mathbf{X} = (x_1^T, \cdots, x_n^T)$ into K groups so that $\mathbf{X}^T = \big(X_1^T, \ldots, X_K^T\big)$ 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 (16), 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

$$\widehat{\mathrm{PR}}(\nu) = -\frac{1}{n}\sum_{k=1}^{K} l_k\big(\widehat{\Sigma}_\nu^{[-k]}, X_k\big). \tag{17}$$

As the an optimal value for the regularization parameter $\kappa_{\max}$, we select ν that minimizes (17),

$$\widehat{\kappa}_{\max} = \inf\left\{\nu|\ \widehat{\mathrm{PR}}(\nu) \le \widehat{\mathrm{PR}}(\nu'),\ \forall \nu' \ge 1\right\}.$$

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

3.2 Properties of the optimal regularization parameter

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

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

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

(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 regularization parameter $\widehat{\delta}$ 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 $\widehat{\delta}$ 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 3. *For a given p,*

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

Proof. Given in Supplemental Section D. □

We resort to numerical methods to demonstrate (P2)–(P4). To this end, we use data sets sampled from zero-mean p-variate Gaussian distributions with the following covariance matrices:

(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 optimal $\widehat{\kappa}_{\max}$ chosen by the proposed selection procedure satisfies the properties (P1)–(P4).

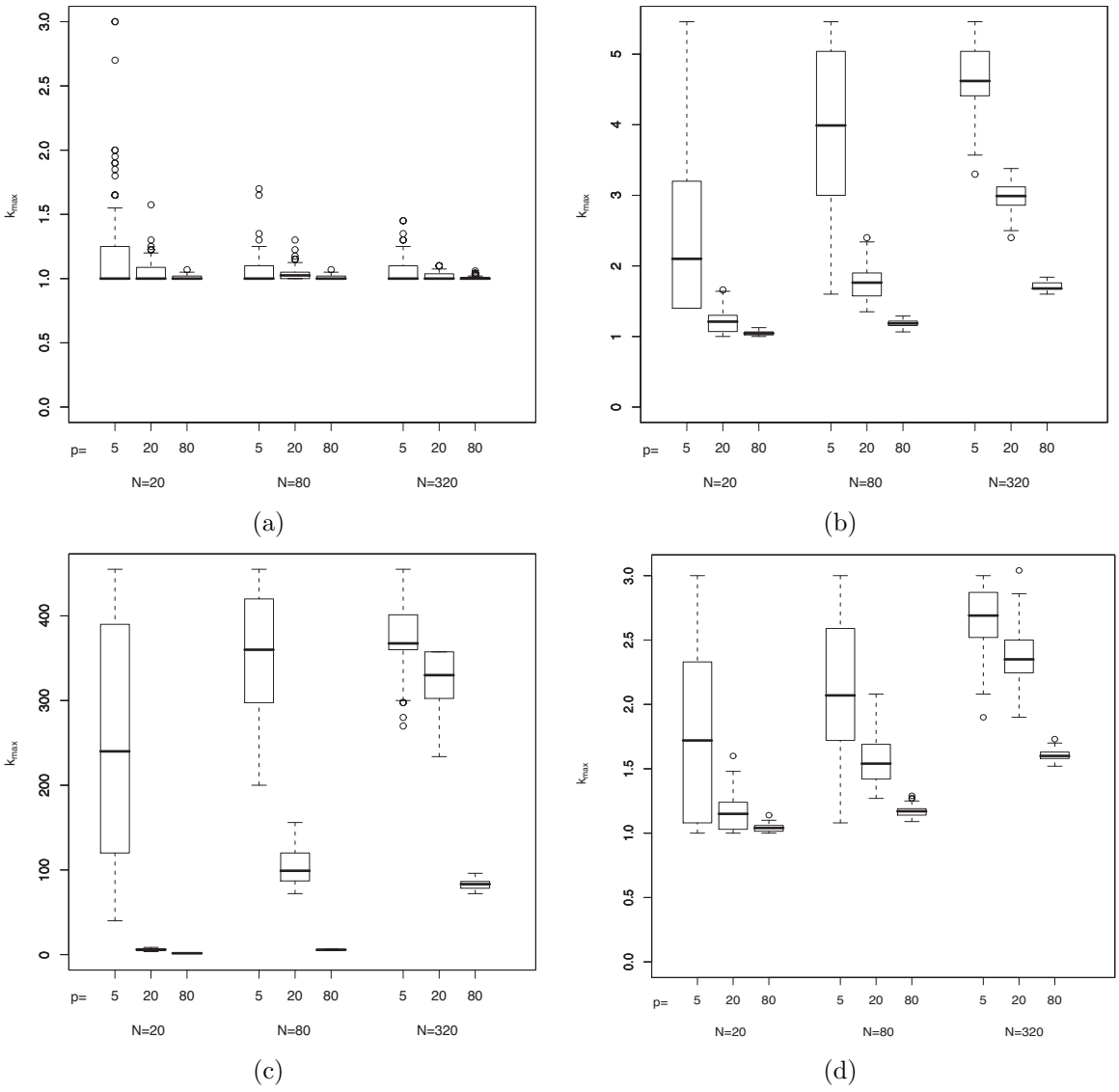


Figure 3: Distribution of $\widehat{\kappa}_{\max}$ for dimensions 5, 20, 80, and for sample sizes 20, 80, 320, with covariance matrices (a) identity (b) diagonal exponentially decreasing, condition number 5, (c) diagonal exponentially decreasing, condition number 400, (d) Toeplitz matrix whose (i, j)th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

4 Bayesian interpretation

Tibshirani (1996) gives a Bayes interpretation for the regularization method 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 same spirit, we can draw parallels for the regularization of covariance matrices by imposing the condition number constraint.

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. This equivalence follows from considering the Lagrangian function (see Gill et al. (1982, Chapter 5)). The condition number-regularized covariance matrix estimation problem (6) can therefore be written in terms of the likelihood of the eigenvalues and the penalty as

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

The above expression allows us to see the condition number-regularized covariance matrix as the Bayes posterior mode under the following prior

$$\pi(\lambda_1, \ldots, \lambda_p) = \exp\left(-g_{\max}\,\frac{\lambda_1}{\lambda_p}\right), \qquad \lambda_1 \geq \cdots \geq \lambda_p > 0 \tag{18}$$

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 as described in the previous sections. First the prior is improper, and hence its properties are similar to a vague prior; but the posterior is always proper.

Proposition 2. *The prior on the eigenvalues implied by the conditioned number constraint*

is improper whereas the posterior yields a proper distribution. More formally,

$$\int_C \pi(\underset{\sim}{\lambda}) d\underset{\sim}{\lambda} = \int_C \exp\left(-g_{\max}\,\frac{\lambda_1}{\lambda_p}\right) d\underset{\sim}{\lambda} = \infty,$$

and

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

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

Proof. Given in Supplemental Section E. □

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. Note that the support of both the prior and the posterior is the entire space of the ordered eigenvalues. So the prior by itself does not give a hard constraint on the condition number. Evaluating the maximum *a posteriori* estimate (MAP) yields an estimator that satisfies the condition number constraint.

A clear picture of the regularization achieved by the prior above and its potential for "eigenvalue 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. 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.*, $\Sigma^{-1} \sim \text{Wishart}(m, cI)$, so that this prior yields a posterior mode which is a linear shrinkage estimator (5) with $\delta = m/(n+m)$. Note however that the coefficients of the combination do not depend of the data X and only on the sample size n and m, the degrees of freedom or shape parameter from the prior. The Ledoit-Wolf estimator (Ledoit and Wolf, 2004b) chooses δ under a data-dependent

optimality – 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). In this prior the eigenvalues are inversely proportional to the determinant $\prod_{i=1}^{p} \lambda_i$ of the the covariance matrix, encouraging shrinkage of the eigenvalues – though unlike the Ledoit-Wolf estimator, the motivation for the reference prior does not stem from obtaining well-conditioned estimators in high-dimensional problems. The posterior mode using this reference prior can be formulated similarly to that of condition number regularization:

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

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

5 Application to two-sample testing

5.1 Intuition

We investigate the use of the condition number-regularized covariance matrix $\widehat{\Sigma}_{\text{cond}}$ for testing a hypothesis about the means in two-sample problems. Consider a test statistic $T(\mathbf{X})$ based on the data matrix $\mathbf{X} = (x_1^T, \ldots, x_n^T)$. Warton (2008, Theorem 5) shows that for test statistics

of the form $T(\mathbf{X}) = g(\Omega_X S^{-1/2})$ and $T_{\text{LS}}(\mathbf{X}) = g(\Omega_X \widehat{\Sigma}_{\text{LS}}^{-1/2})$ that uses the linear shrinkage estimator in place of the sample covariance matrix, where $g(\cdot)$ is some function, Ω_X is some matrix of effects, and S is the sample covariance matrix,

$$T(\mathbf{X}) = g(\Omega_Z \text{diag}(l_1^{-1/2}, \ldots, l_p^{-1/2}) Q^T),$$

and

$$T_{\text{LS}}(\mathbf{X}) = g(\Omega_Z \text{diag}(((1-\delta)l_1 + \delta\gamma)^{-1/2}, \ldots, ((1-\delta)l_p + \delta\gamma)^{-1/2}) Q^T),$$

with $\Omega_Z = \Omega_X Q$ and Q from the spectral decomposition (3) of S. An obvious extension of this result to the condition number-regularized covariance matrix is

$$T_{\text{cond}}(\mathbf{X}) = g(\Omega_Z \text{diag}(\widehat{\lambda}_1^{-1/2}, \ldots, \widehat{\lambda}_p^{-1/2}) Q^T),$$

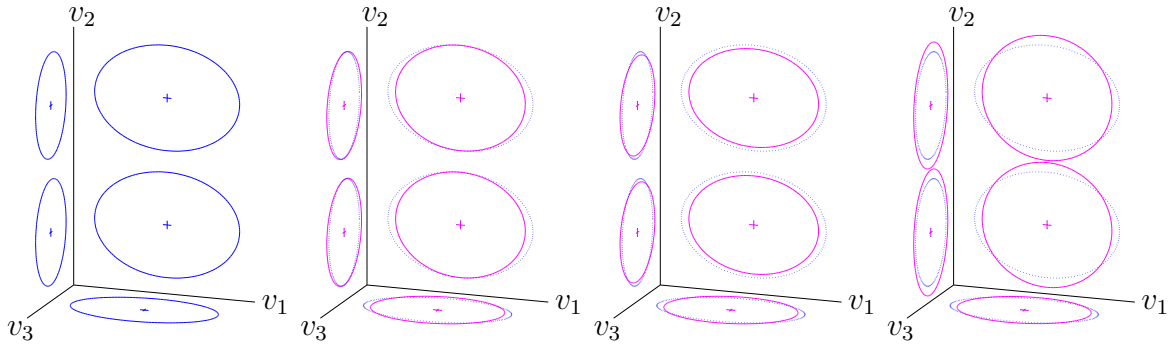
where

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

All of these test statistics are of the same form. The only difference is in the relative weighting imposed to the columns of Ω_Z, the effects expressed in the directions of the eigenvectors of S.

For the regularized test statistics $T_{\text{LS}}(\mathbf{X})$ and $T_{\text{cond}}(\mathbf{X})$, power is expected to increase when effects are expressed along the eigenvectors associated with largest eigenvalues and to decrease otherwise, as they both shrink largest eigenvalues and inflates smallest eigenvalues. They differ mostly in how they treat middle eigenvalues. While $T_{\text{cond}}(\mathbf{X})$ leaves them intact, $T_{\text{LS}}(\mathbf{X})$ shrinks some and inflate the others. Consider three scenarios that contrasts the difference between these test statistics for $p = 3$: when effects are expressed along 1) the eigenvector associated with the largest eigenvalue, 2) the eigenvector associated with the smallest eigenvalue, and 3) the eigenvector associated with the middle eigenvalue. For each of

the scenarios, T_{LS} bears two cases: one that shrinks the middle eigenvalue (case 1), the other that inflates it (case 2). The test statistics can be understood to evaluate distance between two ellipsoids whose shapes are determined by the estimate of the common covariance matrix and only centers differ. In scenario 1, both regularization schemes render the ellipsoids further apart, improving the power compared to that of $T(\mathbf{X})$. In scenario 2, $T_{\text{cond}}(\mathbf{X})$ does not shrink the ellipsoid in the direction of the effect, thus the power does not change. On the other hand, case 1 for $T_{\text{LS}}(\mathbf{X})$ shrinks the ellipsoid while case 2 inflates. Hence the power may be improved or reduced. In scenario 3, the ellipsoids are brought closer to each other by both regularization scheme, and the power is reduced. Scenario 2, which reveals the difference between $T_{\text{cond}}(\mathbf{X})$ and T_{LS} clearly, is illustrated in Figure 4. Illustration of the other scenarios can be found in Supplemental Section G.



(a) $T(\mathbf{X})$ (b) $T_{\text{cond}}(\mathbf{X})$ (c) $T_{\text{LS}}(\mathbf{X})$, case 1 (d) $T_{\text{LS}}(\mathbf{X})$, case 2

Figure 4: Schematic diagram illustrating the effects of regularization on the two-sample test statistic $T(\mathbf{X})$ in three dimensions. Each of the two ellipsoids are visualized by three ellipses projected onto the planes made of pairs of eigenvectors. Means differ along the eigenvector (v_2) associated with the middle eigenvalue.

5.2 Power simulation

We demonstrate power properties of $T_{\text{cond}}(\mathbf{X})$ and $T_{\text{LS}}(\mathbf{X})$ by numerical simulations. In each simulation we generate two sets of independent samples, of respective sizes n_1 and n_2,

from the p-variate Gaussian distribution having a common covariance matrix Σ, with mean vectors $\mu_1 = 0$ and μ_2 shifted from 0. We consider three types of shifts:

- (first) Along the eigenvector of Σ associated with the largest eigenvalue by $a\sqrt{p\lambda_1}$ units
- (last) Along the eigenvector of Σ associated with the smallest eigenvalue by $a\sqrt{p\lambda_p}$ units
- (upper) Along the eigenvectors associated with the largest half eigenvalues: $a\sqrt{2\lambda_i}$ units along the i-th eigenvector of Σ, $i = 1, \ldots, \lfloor p/2 \rfloor$
- (lower) Along the eigenvectors associated with the smallest half eigenvalues: $a\sqrt{2\lambda_i}$ units along the i-th eigenvector of Σ, $i = \lfloor p/2 \rfloor + 1, \ldots, p$
- (all) Along all eigenvectors: $a\sqrt{\lambda_i}$ units along the i-th eigenvector of Σ, $i = 1, \ldots, p$.

For each of the shifts, we test the null hypothesis $\mathcal{H}_0 : \mu_1 = \mu_2$ using the regularized versions (*i.e.*, $T_{\mathrm{LS}}(\mathbf{X})$ and $T_{\mathrm{cond}}(\mathbf{X})$)) of $T(\mathbf{X}) = -2\log(\mathbf{\Lambda})$, where $\mathbf{\Lambda}$ is Wilks' lambda statistic (Mardia et al., 1980), $\mathbf{\Lambda} = |(n_1+n_2)\widehat{\Sigma}|/|(n_1+n_2)\widehat{\Sigma} + B| = 1/|I + \widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}/(n_1+n_2)|$, where B is the between-groups matrix. As a reference, we also employed $T_{-}(\mathbf{X})$, in which the Moore-Penrose inverse is used in place of $\widehat{\Sigma}^{-1}$ in $T(\mathbf{X})$.

We use $n_1 = n_2 = 10$, $p \in \{5, 10, 20\}$, $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho \in \{0.3, 0.5\}$ (*i.e.*, AR(1) structure with unit variance), and $a = 0.3$ to obtain intermediate power for most of the simulation scenarios. We use the Ledoit-Wolf optimality in choosing the regularization parameter δ for $T_{\mathrm{LS}}(\mathbf{X})$ and the predictive risk (16) in selecting $\kappa_{\max}$ for $T_{\mathrm{cond}}(\mathbf{X})$). We use permutation tests where the P-values are computed from 999 permutations of group membership labels for the $n_1 + n_2$ observations. The two samples are adjusted to have equal sample means, following the recipe of Warton (2008). We estimate power of $T_{\mathrm{LS}}(\mathbf{X})$, $T_{\mathrm{cond}}(\mathbf{X})$ and $T_{-}(\mathbf{X})$ from 400 data sets for each p and ρ by counting the number of data sets which reject $\mathcal{H}_0$, at the significance levels .10 and .05, respectively.

The results, which can be found in Supplemental Section H, show that $T_{\text{LS}}(\mathbf{X})$ and $T_{\text{cond}}(\mathbf{X})$ behave similarly. This is expected from the regularization pattern discussed in the previous section. In particular,

- Power of $T_{\text{LS}}(\mathbf{X})$ and $T_{\text{cond}}(\mathbf{X})$ tends to increase as p increases. This behavior is consistent with that reported by Warton (2008).

- Power of $T_{-}(\mathbf{X})$ tends to be smallest at $p = 10$. This behavior is also reported by Warton (2008).

- Power is relatively high when the effects are expressed along the eigenvectors associated with large eigenvalues ("first" and "upper"). $T_{\text{LS}}(\mathbf{X})$ seems to have higher power when the effect is one-dimensional ("first"), whereas $T_{\text{cond}}(\mathbf{X})$ seems to have higher power when the effect is multidimensional ("upper").

- Power is relatively low when the effects are expressed along the eigenvectors associated with small eigenvalues ("last" and "lower"). Sometimes $T_{\text{LS}}(\mathbf{X})$ and $T_{\text{cond}}(\mathbf{X})$ are beaten by $T_{-}(\mathbf{X})$. $T_{\text{LS}}(\mathbf{X})$ seems to have higher power when the effect is one-dimensional ("last"), whereas $T_{\text{cond}}(\mathbf{X})$ seems to have higher power when the effect is multidimensional ("lower").

- When the effect is expressed in all dimensions by a small amount ("all"), the power is intermediate between "upper" and "lower," while $T_{\text{cond}}(\mathbf{X})$ seems better.

The advantage of $T_{\text{cond}}(\mathbf{X})$ over $T_{\text{LS}}(\mathbf{X})$ for the effects expressed in many directions can be understood by virtue of Figure 4. When the effects involve eigenvectors associated with middle eigenvalues, power increases if the middle eigenvalues are shrunk and decreases if inflated. $T_{\text{LS}}(\mathbf{X})$ shrinks some of these eigenvalues and inflates the other some. Choosing the border line is rather a gamble, and it appears that this gambling is not as good as having them remain intact, which is what $T_{\text{cond}}(\mathbf{X})$ does.

6 Application to robust portfolio selection

This section illustrates the merits of the proposed regularization scheme in portfolio optimization. We consider a portfolio rebalancing strategy based on robust mean-variance portfolio selection. A portfolio refers to a collection of risky assets held by an institution or an individual. Over the holding period, the return on the portfolio is the weighted average of the returns on the individual assets that constitutes the portfolio, where the weight of each asset corresponds to its proportion in the portfolio. The portfolio optimization problem concerns the weights that maximizes the return on the portfolio. However, since the asset returns are stochastic, a portfolio always carries a risk of loss. The mean-variance portfolio (MVP) theory (Markowitz, 1952) quantifies the risk of a portfolio with the standard deviation of its returns. Estimation of the covariance of asset returns thus become crucial in this setting. The other important component of the MVP theory is the expected return on the portfolio. Unfortunately, it is extremely difficult to estimate the expected asset returns (Luenberger, 1998; Merton, 1980). Since the focus of this paper lies in estimating covariance matrices and not in expected returns, we cope with this difficulty by requiring the portfolio to be robust to uncertainty in estimated expected asset returns. Indeed, the problem of robust portfolio optimization is gaining popularity in financial literature (Ceria and Stubbs, 2006; Goldfarb and Iyengar, 2003; Tütüncü and Koenig, 2004).

We use the condition number-regularized covariance matrix, and two existing ones, namely, a linear shrinkage estimator and the sample covariance matrix, in constructing a robust mean-variance portfolio. We compare their performance over a period of more than 14 years.

6.1 Robust mean-variance portfolio rebalancing

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

In the MVP theory, the expected return of a portfolio w is denoted by $\mu^T w$, and the risk by the standard deviation $(w^T \Sigma w)^{1/2}$. Then the mean-variance portfolio selection problem maximizes a quadratic utility function as follows.

$$\begin{array}{ll} \text{maximize} & \mu^T w - \frac{\gamma}{2} w^T \Sigma w \\ \text{subject to} & \mathbf{1}^T w = 1, \end{array} \tag{19}$$

where γ is the parameter of relative risk aversion. This is a simple quadratic program that has an analytic solution. In practice, both μ and Σ should be estimated. As noted, estimation of μ is much harder than that of Σ, and estimation error in μ has larger impact on deviation of portfolio weights from its optimal value than that in Σ (DeMiguel and Nogales, 2009; Merton, 1980). In other words, plugging in a (regularized) sample covariance matrix estimated from historical returns in place of Σ may suffice, but this is hardly true for μ.

As a means to introduce robustness to error in estimating μ, we consider a MVP that utilizes a "worst-case" vector of expected returns (Ceria and Stubbs, 2006). Suppose we are uncertain the expected return μ but know that it belongs to a set $\mathcal{E}$ with some confidence. We

would like to maximize our utility for the worst-case expected portfolio return $\min_{\mu\in\mathcal{E}} \mu^T w$, *i.e.*,

$$\begin{array}{ll} \text{maximize} & \min_{\mu\in\mathcal{E}} \mu^T w - \frac{\gamma}{2} w^T \Sigma w \\ \text{subject to} & \mathbf{1}^T w = 1, \end{array} \tag{20}$$

instead of (19). If we use the (historical) sample mean $\overline{r}$ as an estimator of μ, we can employ a $100(1-\alpha)\%$ confidence ellipsoid for μ in place of $\mathcal{E}$. Assume r follows a p-variate Gaussian distribution and let N_{estim} be the size of the estimation horizon, *i.e.*, the number of past data points used to estimate μ and Σ. Then $\overline{r} \sim \mathcal{N}(\mu, \frac{1}{N_{\text{estim}}}\Sigma)$ and

$$\mathcal{E} = \{\mu : (\mu - \overline{r})^T \Sigma^{-1} (\mu - \overline{r}) \leq \chi^2_{1-\alpha}(p)/N_{\text{estim}}\},$$

where $\chi^2_{1-\alpha}(p)$ is the $100(1-\alpha)\%$ quantile of the chi-square distribution with p degrees of freedom. For this choice of $\mathcal{E}$, it is easy to see that $\min_{\mu\in\mathcal{E}} \mu^T w = \overline{r}^T w - (\chi^2_{1-\alpha}(p)/N_{\text{estim}})^{1/2}\sqrt{w^T \Sigma w}$. Now (20) becomes

$$\begin{array}{ll} \text{maximize} & \overline{r}^T w - (\chi^2_{1-\alpha}(p)/N_{\text{estim}})^{1/2}\sqrt{w^T \Sigma w} - \frac{\gamma}{2} w^T \Sigma w \\ \text{subject to} & \mathbf{1}^T w = 1, \end{array} \tag{21}$$

which is a convex optimization problem. We call (21) the robust mean-variance portfolio selection problem.

The portfolio selection problem 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 robust mean-variance portfolio rebalancing (rMVR) strategy. Let $r^{(t)} = (r_1^{(t)}, \ldots, r_p^{(t)}) \in \mathbb{R}^p$, $t = 1, \ldots, N_{\text{tot}}$, be the realized returns of assets at time t. (The time unit can be a day, a week, or a month.) We consider periodic mean-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 robust mean-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, as defined above. 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 for this holding period is not taken into account.) We therefore have a series of portfolios $w^{(j)}$ that solves

$$\begin{array}{ll} \text{maximize} & \overline{r}^{(j)T} w - (\chi^2_{1-\alpha}(p)/N_{\text{estim}})^{1/2} \sqrt{w^T \hat{\Sigma}^{(j)} w} - \frac{\gamma}{2} w^T \hat{\Sigma}^j w \\ \text{subject to} & \mathbf{1}^T w = 1, \end{array}$$

over the holding periods of $[N_{\text{estim}} + 1 + (j-1)L, N_{\text{estim}} + jL],\ j = 1, \ldots, K$. Here $\overline{r}^{(j)}$ is the sample mean, and $\hat{\Sigma}^{(j)}$ is the covariance matrix, of the asset returns estimated from those over the jth holding period.

6.2 Empirical out-of-sample performance

In this empirical study, we use the 30 stocks that constituted the Dow Jones Industrial Average over the period from February 1994 to July 2008. (Supplemental Section I.1 lists 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 whole period considered in our numerical study is from the first trading date in March 2, 1992 to July 14, 2008. (This period consists of 4125 trading days.) The time unit used in our study is 5 consecutive trading days, so we consider weekly returns. We take

$$N_{\text{tot}} = 825, \qquad L = 25, \qquad N_{\text{estim}} = 15, 30, 45, 60.$$

To estimate the covariance matrices, we use the last N_{estim} weekly returns of the constituents

of the Dow Jones Industrial Average. Roughly in every half year, we rebalance the portfolio, using a covariance estimate with past roughly two-year weekly return data. (Supplemental Section I.2 shows the periods determined by the choice of the parameters.) The trading period corresponds to $K = 29$ holding periods, which span the dates from February 18, 1994 to July 14, 2008.

In the sequel, we compare the rMVR strategy where the covariance matrices are estimated using the condition number-regularization scheme with the rMVR strategies using the sample covariance matrix and the linear shrinkage estimator under the Ledoit-Wolf optimality.

Performance metrics

We use the following quantities in assessing the performance of the rMVR strategies.

- *Realized return.* The realized return of the portfolio over the trading period.
- *Realized risk.* The realized risk (return standard deviation) of the portfolio over the trading period.
- *Realized utility.* The realized value of the quadratic utility function of the portfolio over the trading period.
- *Turnover.* Total amount of new portfolio assets purchased or sold over the trading period.
- *Normalized wealth growth.* Accumulated wealth yielded by the portfolio over the trading period when the initial budget is normalized to one, taking the transaction cost into account.

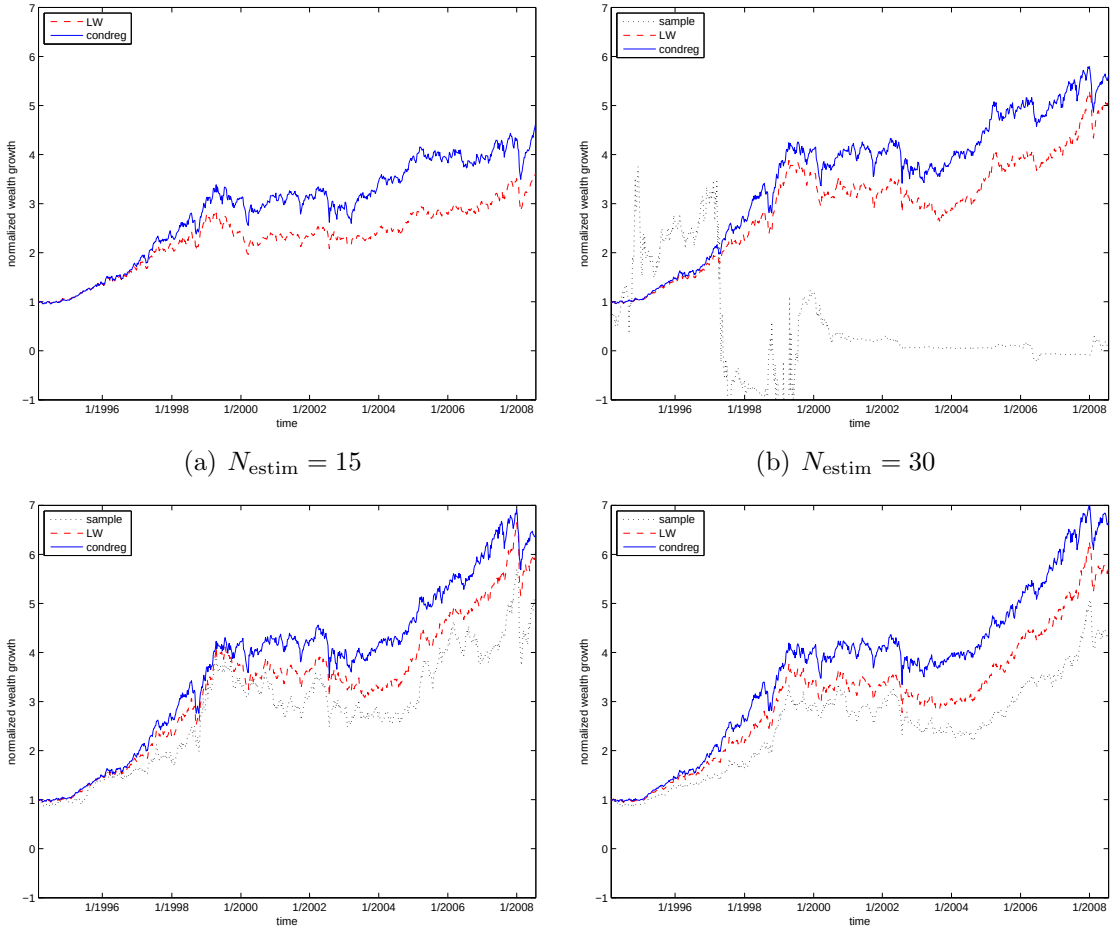
For precise formulae of these metrics, refer to Supplemental Section I.3. We assume that the transaction costs are the same for the 30 stocks and set them to 30 basis points (bps). We choose the relative risk aversion parameter $\gamma = 25$ and use a 95% ($\alpha = 0.05$) confidence ellipsoid for the expected return.

Comparison results

Figure 5 shows the normalized wealth growth over the trading horizon from February 18, 1994 through July 14, 2008. (The sample covariance matrix failed in solving (21) for $N_{\text{estim}} = 15$ because of its singularity and hence omitted.) The rMVR strategy using the condition number-regularized covariance matrix outperforms significantly the rMVR strategy using the sample covariance matrix or the linear shrinkage estimator.

Rather surprisingly, there is no significant difference between the condition number regularization and the linear shrinkage in terms of MVP metrics. Supplemental Section I.4 summarizes the realized utility, the realized return, and the realized risk of each estimator respectively averaged over the trading period. For all values of N_{estim}, the average differences of the metrics between the two regularization schemes are roughly within two standard errors of those. On the other hand, both the condition number-regularized matrix and the linear shrinkage estimator outperforms the sample covariance matrix especially when the estimation horizon size N_{estim} is less than or moderately greater than the number of stocks in the portfolio.

Then what makes the difference in wealth growth? It is turnover of the portfolio. In Appendix I.5 we can see that rMVR using the condition number-regularized covariance matrix gives a far lower turnover and thus more stable weights than rMVR using the linear shrinkage estimator or the sample covariance matrix. A lower turnover implies less transaction costs, thereby contributing to the higher wealth growth. The stability of the rMVR portfolio using the proposed estimator is an interesting phenomenon that calls for further research, as no explicit effort was made to limit turnover.



(a) $N_{\text{estim}} = 15$ (b) $N_{\text{estim}} = 30$

(c) $N_{\text{estim}} = 45$ (d) $N_{\text{estim}} = 60$

Figure 5: Robust mean-variance rebalancing results for various estimation horizon sizes over the trading period from February 18, 1994 through July 14, 2008. `sample`=sample covariance matrix, `LW`=linear shrinkage using the Ledoit-Wolf optimality, `condreg`=condition number regularization.

7 Conclusions

In this paper we considered regularized covariance estimation by imposing a constraint on the condition number in the Gaussian likelihood formulation. Regularization schemes that have been proposed in the literature for high-dimensional covariance estimation do not *directly* target the issue of invertibility and numerical stability of the estimate. We have emphasized the importance of a numerically well-conditioned estimator of covariance matrices, especially in practical applications such as two-sample testing and portfolio selection. A consequence of this emphasis on numerical stability is the condition number-regularized maximum likelihood estimator. We have shown that this regularization scheme 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 explored the theoretical properties of the proposed regularization and shown that the resulting estimator asymptotically dominates the sample covariance estimator with respect to the entropy loss under a mild assumption. We have also provided a cross-validated parameter selection procedure. The cross-validated estimator demonstrates a desired performance in two-sample testing compared with other commonly used estimators of covariance matrices. When applied to a real-world wealth management problem, our regularization scheme performs very well, supporting its usefulness in a variety of applications where a well-conditioned estimate of the covariance matrix is often required.

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.

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

Ceria, S. and R. A. Stubbs (2006, July). Incorporating estimation errors into portfolio selection: Robust portfolio construction. *Journal of Asset Management 7*, 109–127.

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

DeMiguel, V. and F. J. Nogales (2009). Portfolio selection with robust estimation. *Operations Research 57*(3), 560–577.

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.

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

Gill, P. E., W. Murray, and M. H. Wright (1982, February). *Practical Optimization*. Academic Press.

Goldfarb, D. and G. Iyengar (2003). Robust portfolio selection problems. *Mathematics of Operations Research 28*(1), 1–38.

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

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.

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.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1980, February). *Multivariate Analysis*. Academic Press.

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.

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

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.

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

Tütüncü, R. H. and M. Koenig (2004, November). Robust asset allocation. *Annals of Operations Research 132*(1), 157–187.

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.

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

Supplemental Materials

A Proof of Theorem 1

A.1 Uniqueness of the solution to (13)

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, $1/\big(\kappa_{\max} l_p\big) < 1/l_1$, *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.

A.2 An algorithm for solving (13)

Without loss of generality, we assume that $\kappa_{\max} < l_1/l_p = \mathrm{cond}(S)$. As discussed above, the function $J_{\kappa_{\max}}(u) \triangleq \sum_{i=1}^{p} J^{(i)}_{\kappa_{\max}}(u)$ 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 (14):

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

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

$$1/l_\alpha < u_{\alpha,\beta} \leq 1/l_{\alpha+1} \tag{22}$$

and

$$1/l_{\beta-1} \le \kappa_{\max} u_{\alpha,\beta} < 1/l_\beta. \tag{23}$$

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 (22) and (23), hence $u^\star$, in $O(p)$ operations. Algorithm 1 describes the procedure. This procedure was first considered by Won and Kim (2006) and is elaborated in this paper.

B Proof of Proposition 1

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

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

and

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

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

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

and

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

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

1. $1/l_\alpha < u^\star(\nu_0) < 1/l_{\alpha+1}$ and $1/l_{\beta-1} < v^\star(\nu_0) < 1/l_\beta$.
   We can find $\nu > \nu_0$ such that

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

   and

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

   Therefore,

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

Algorithm 1 Solution method to the optimization problem (13)

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

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,

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

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.

C 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 (3). Then the solution to (9) 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

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

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

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

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

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

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 $S_0\triangleq A^TSA$ has the same distribution as the sample covariance matrix observed from a p-variate Gaussian distribution with the identity covariance matrix. From the operational definition of the largest eigenvalue l_1 of S, we obtain

$$l_1=\max_{v\neq0}\frac{v^TSv}{v^Tv}=\max_{w\neq0}\frac{w^TS_0w}{w^T\Lambda^{-1}w},$$

where $w=A^Tv$. Furthermore, since for any $w\neq0$,

$$\lambda_1^{-1}=\min_{w\neq0}\frac{w^T\Lambda^{-1}w}{w^Tw}\le\frac{w^T\Lambda^{-1}w}{w^Tw},$$

we have

$$l_1\le\lambda_1\max_{w\neq0}\frac{w^TS_0w}{w^Tw}=\lambda_1e_1,\tag{25}$$

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

$$l_p\ge\lambda_pe_p,\tag{26}$$

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

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

The combination of (25)–(27) leads to

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

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 Section A.2, $u^\star$ 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.

D Proof of Theorem 3

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 $\nu \geq l_1^{[-k]}/l_p^{[-k]} = \mathrm{cond}(S^{[-k]})$,

$$\widehat{\kappa}_{\max} \leq \max_{k=1,\ldots,K} l_1^{[-k]}/l_p^{[-k]}. \tag{28}$$

The right hand side of (28) 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 (9). We denote the limit of $\widehat{\Sigma}_\nu^{[-k]}$ by $\tilde{\Sigma}_\nu$. For the spectral decomposition of Σ

$$\Sigma = R\ \mathrm{diag}\big(\lambda_1, \ldots, \lambda_p\big) R^T, \tag{29}$$

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

$$\tilde{\Sigma}_\nu = R\ \mathrm{diag}\big(\psi_1(\nu), \ldots, \psi_p(\nu)\big)\ R^T, \tag{30}$$

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,

$$-\frac{1}{n}\frac{\partial l_k\big(\widehat{\Sigma}_\nu^{[-k]}, X_k\big)}{\partial \nu} \to c_k \,\mathbf{Tr}\left\{\frac{\partial \tilde{\Sigma}_\nu^{-1}}{\partial \nu}\Big(\Sigma - \tilde{\Sigma}_\nu\Big)\right\}, \quad \text{almost surely.} \tag{31}$$

We now study (31). First, if $\nu \geq \kappa$, then $\tilde{\Sigma}_\nu = \Sigma$, and the RHS of (31) degenerates to 0. Now we consider the non-trivial case that $\nu < \kappa$. From (30),

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

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 (29) and (30),

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

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 (31) 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 Proposition 2

We are given that

$$\pi(\lambda_1, \ldots, \lambda_p) = \exp\left(-g_{\max}\,\frac{\lambda_1}{\lambda_p}\right) \qquad \lambda_1 \geq \cdots \geq \lambda_p > 0.$$

Now

$$\int_C \pi(\lambda_1, \ldots, \lambda_p)d\underset{\sim}{\lambda} = \int_C \exp\left(-g_{\max}\,\frac{\lambda_1}{\lambda_p}\right)d\underset{\sim}{\lambda},$$

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

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

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

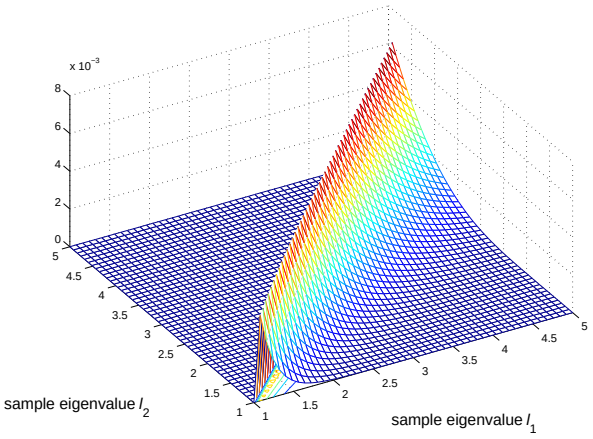
To prove that the posterior yields a proper distribution we proceed as follows:

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

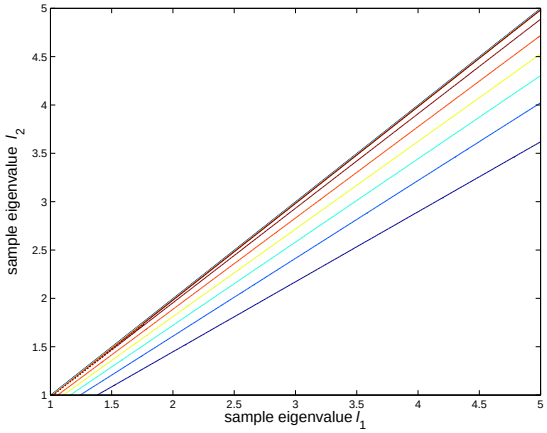
The above integrand is the density of the inverse gamma distribution and therefore the corresponding integral above has a finite normalizing constant and thus yielding a proper posterior.

F Comparison of Bayesian prior densities

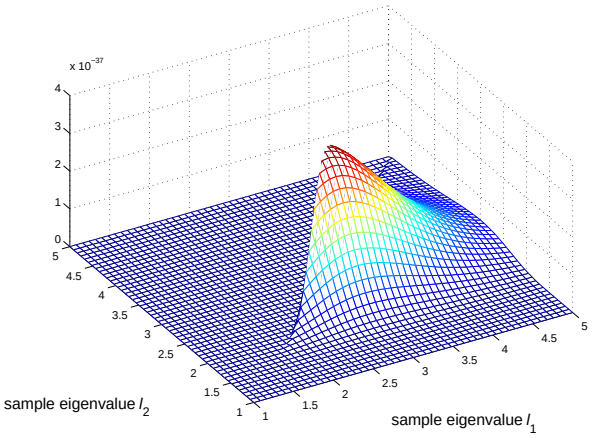
Comparison of various prior densities for eigenvalue shrinkage ($p = 2$). (a) Three-dimensional, (b) contour view of the prior density (18). (c) Three-dimensional, (d) contour view of the prior density induced by the inverse Wishart distribution. (a) Three-dimensional, (b) contour view of the reference prior density due to Yang and Berger (1994).


x 10^-3
sample eigenvalue l_2
sample eigenvalue l_1

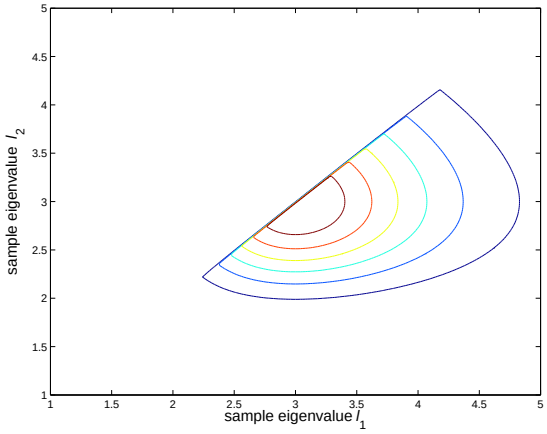
(a)


sample eigenvalue l_2
sample eigenvalue l_1

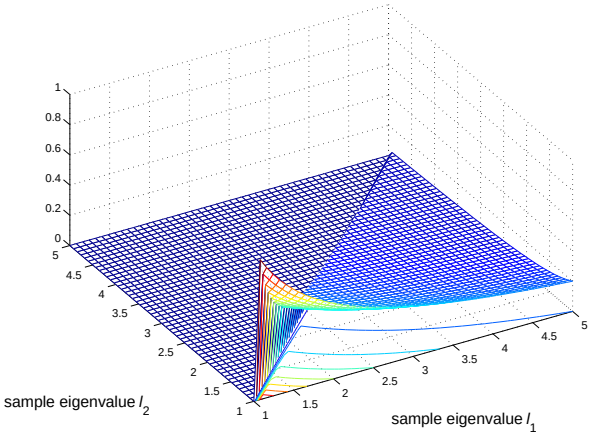
(b)


x 10^-37
sample eigenvalue l_2
sample eigenvalue l_1

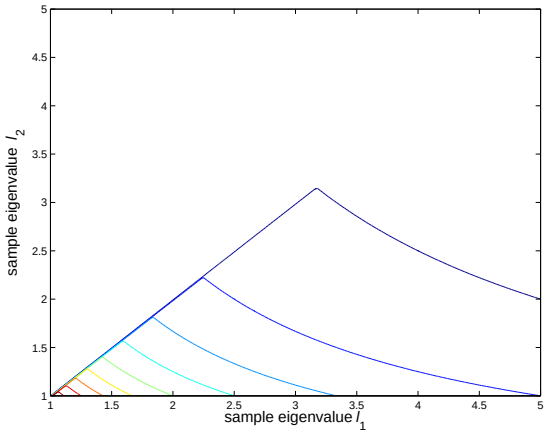
(c)


sample eigenvalue l_2
sample eigenvalue l_1

(d)


sample eigenvalue l_2
sample eigenvalue l_1

(e)

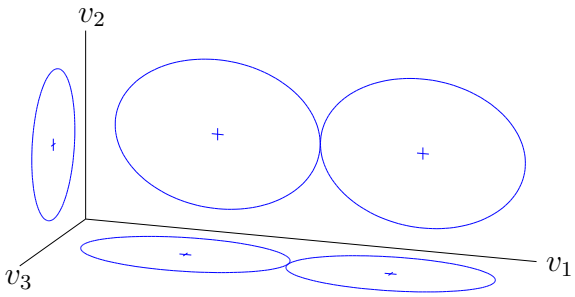

sample eigenvalue l_2
sample eigenvalue l_1

(f)

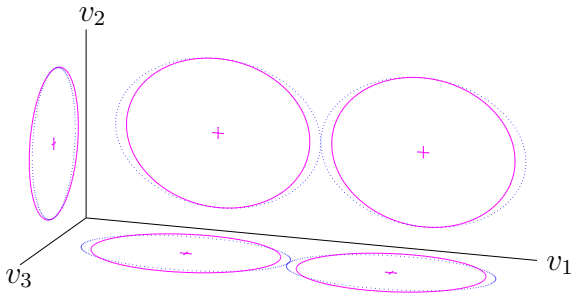
G Illustration of the effects of regularization in two-sample testing

Schematic diagrams illustrating the effects of regularization on the two-sample test statistic $T(\mathbf{X})$ in three dimensions. Each of the two ellipsoids are visualized by three ellipses projected onto the planes made of pairs of eigenvectors.

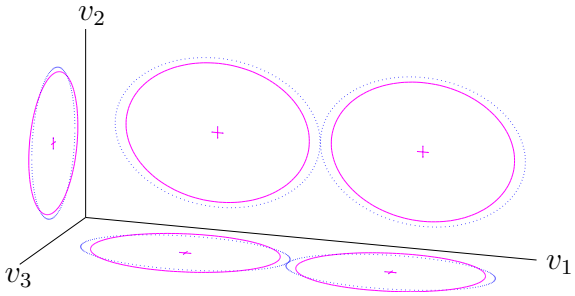
G.1 Means differ along the eigenvector (v_1) associated with the largest eigenvalue



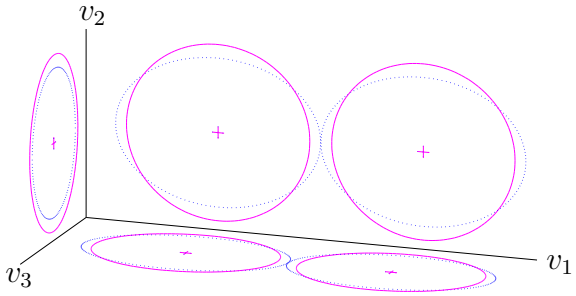
(a) $T(\mathbf{X})$



(b) $T_{\text{cond}}(\mathbf{X})$

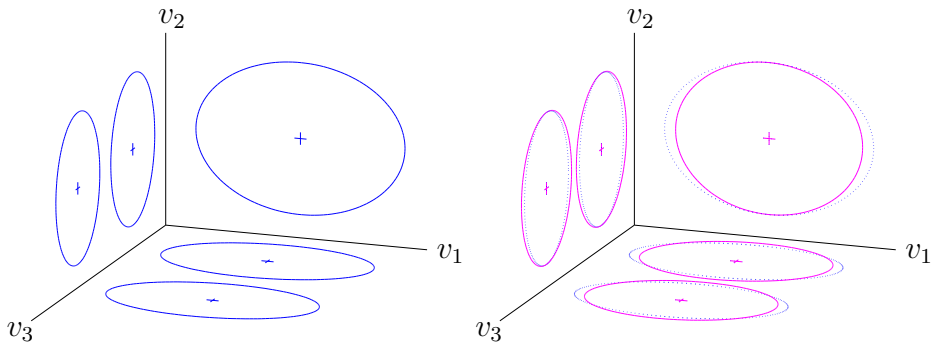


(c) $T_{\text{LS}}(\mathbf{X})$, case 1

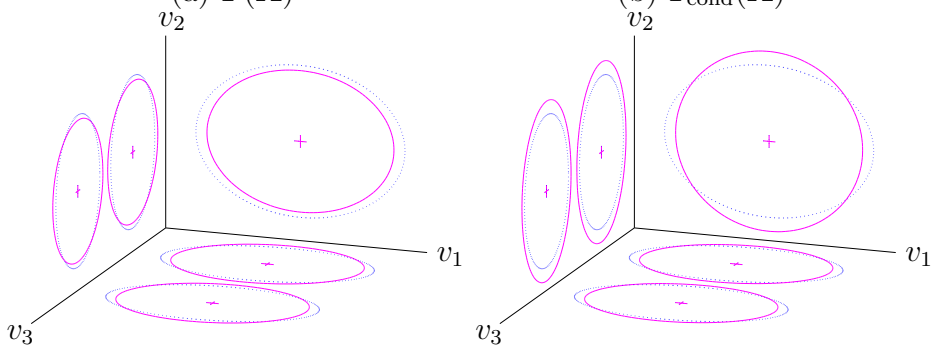


(d) $T_{\text{LS}}(\mathbf{X})$, case 2

G.2 Means differ along the eigenvector (v_3) associated with the smallest eigenvalue



(a) $T(\mathbf{X})$ (b) $T_{\text{cond}}(\mathbf{X})$



(c) $T_{\text{LS}}(\mathbf{X})$, case 1 (d) $T_{\text{LS}}(\mathbf{X})$, case 2

H Power simulation results

Power of different test statistics in which the sample estimate of the covariance matrix is regularized. For each simulation scenario and level, the test statistic that achieves the highest power is highlighted.

ρ	shift	p	level=0.10			level=0.05		
			$T_{\mathrm{LS}}(\mathbf{X})$	$T_{\mathrm{cond}}(\mathbf{X})$	$T_{-}(\mathbf{X})$	$T_{\mathrm{LS}}(\mathbf{X})$	$T_{\mathrm{cond}}(\mathbf{X})$	$T_{-}(\mathbf{X})$
0.3	first	5	0.3500	**0.3575**	0.3175	**0.2275**	0.2200	0.2050
		10	**0.5200**	0.5125	0.3550	**0.3875**	0.3675	0.2375
		20	**0.7100**	0.7000	0.4000	**0.5850**	0.5700	0.3175
	last	5	0.1550	0.1475	**0.1950**	0.0775	0.0750	**0.1000**
		10	**0.2250**	0.2225	0.1800	**0.1300**	0.1275	0.1000
		20	**0.2625**	0.2475	0.1775	**0.1725**	0.1475	0.1000
	upper	5	0.2500	**0.2550**	0.2175	**0.1725**	0.1700	0.1300
		10	0.3625	**0.3900**	0.2200	0.2475	**0.2725**	0.1000
		20	0.4575	**0.5175**	0.2375	0.3450	**0.4025**	0.1650
	lower	5	0.1875	**0.2100**	0.2075	0.1200	**0.1375**	0.1050
		10	0.2275	**0.2600**	0.1400	0.1600	**0.1675**	0.0700
		20	0.2925	**0.3000**	0.1925	0.1800	**0.1925**	0.1200
	all	5	**0.2575**	0.2550	0.2175	**0.1650**	0.1550	0.1025
		10	0.2650	**0.2950**	0.1725	0.1700	**0.1975**	0.1100
		20	0.4075	**0.4275**	0.2425	0.2750	**0.3200**	0.1450
0.5	first	5	**0.5050**	0.4900	0.5025	**0.3525**	0.3450	0.3500
		10	**0.7525**	0.6675	0.5475	**0.6100**	0.5000	0.3975
		20	**0.9175**	0.8625	0.5675	**0.8500**	0.7475	0.4925
	last	5	**0.1925**	0.1600	0.1775	**0.0925**	**0.0925**	0.0850
		10	0.1475	0.1500	**0.1575**	0.0825	**0.0875**	0.0775
		20	0.1725	0.1300	**0.2050**	**0.1075**	0.0750	**0.1075**
	upper	5	**0.2800**	0.2625	0.2700	**0.1800**	0.1750	0.1650
		10	0.3275	**0.3425**	0.2050	**0.2074**	0.2050	0.1350
		20	0.4375	**0.5300**	0.2675	0.3175	**0.3925**	0.2000
	lower	5	0.1625	0.1600	**0.1850**	0.0650	**0.0750**	**0.0750**
		10	0.1775	**0.1825**	0.1200	0.1100	**0.1200**	0.0550
		20	**0.2125**	0.2075	0.1450	0.1125	**0.1350**	0.0850
	all	5	0.1825	**0.2025**	0.1875	0.0925	**0.1050**	0.0850
		10	0.2125	**0.2500**	0.1525	0.1325	**0.1450**	0.0950
		20	0.3325	**0.3575**	0.2350	0.2000	**0.2675**	0.1475

I Empirical robust mean-variance rebalancing study (Section 6.2)

I.1 List of Dow Jones stocks

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 the Sharpe ratio (SR) are annualized.

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

I.2 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		

I.3 Performance metrics

We use the following quantities in assessing the performance of the rMVR 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

$$\overline{r}^{(j)}(w) = \frac{1}{L} \sum_{t=N_{\text{estim}}+1+(j-1)L}^{N_{\text{estim}}+jL} r^{(t)T} w^{(j)}.$$

- *Realized risk.* 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

$$\sigma^{(j)}(w) = \sqrt{w^{(j)T} \Sigma^{(j)}_{\text{sample}} w^{(j)}},$$

  where $\Sigma^{(j)}_{\text{sample}}$ is the sample covariance matrix of the asset returns over the period.
- *Realized utility.* The realized utility of a portfolio w over the period $[N_{\text{estim}} + 1 + (j-1)L, N_{\text{estim}} + jL]$ is given by

$$u^{(j)}(w) = \overline{r}^{(j)}(w) - \frac{\gamma}{2}(\sigma^{(j)}(w))^2,$$

  where γ is the relative risk aversion parameter in (21).

- *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

$$\mathrm{TO}(j) = \sum_{i=1}^{p} \left| w_i^{(j)} - \left(\prod_{t=N_{\mathrm{estim}}+1+(j-1)L}^{N_{\mathrm{estim}}+jL} r_i^{(t)} \right) w_i^{(j-1)} \right|.$$

  For the first period, we take $w^{(0)} = 0$, *i.e.*, the initial holdings of the assets are zero.
- *Normalized wealth growth.* 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 budget is normalized to one, the normalized wealth grows according to the recursion

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

  for $t = N_{\mathrm{estim}}, \ldots, N_{\mathrm{estim}} + KL$, with the initial wealth $W(N_{\mathrm{estim}}) = 1$. Here

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

  and

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

  is the transaction cost due to the rebalancing if the cost to buy or sell one share of stock i is η_i.

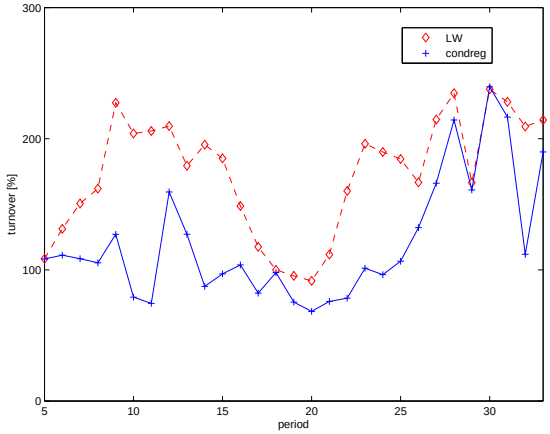
I.4 Mean-variance portfolio theoretic performance metrics

Realized utility, realized return, and realized risk based on different regularization schemes for the covariance matrices are reported. `sample`=sample covariance matrix, `LW`=linear shrinkage using the Ledoit-Wolf optimality, `condreg`=condition number regularization. Each entry is the mean (standard error) of the corresponding metric over the trading period (29 holding periods) from March 1992 through July 2008. All values are annualized.

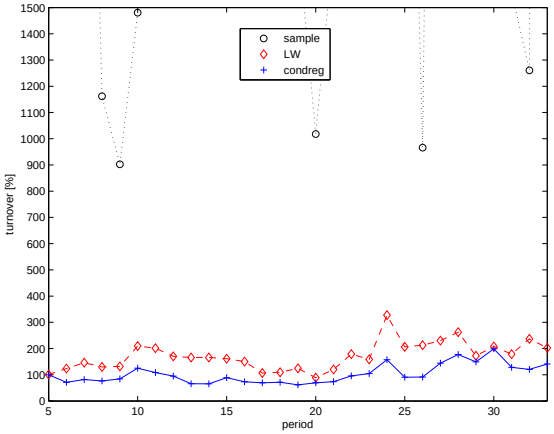
covariance regularization scheme	utility	return [%]	risk [%]
		$N_{\text{estim}} = 15$	
`sample`	–	–	–
`LW`	-0.143 (0.0390)	0.116 (0.0306)	0.138 (0.00732)
`condreg`	-0.168 (0.0507)	0.135 (0.0323)	0.148(0.00889)
`condreg - LW`	-0.0245 (0.0234)	0.0184 (0.0158)	0.00950 (0.00429)
		$N_{\text{estim}} = 30$	
`sample`	-65.4 (29.3)	-0.326 (0.446)	1.41 (0.340)
`LW`	-0.143 (0.0464)	0.143 (0.0371)	0.145 (0.00819)
`condreg`	-0.138 (0.0526)	0.144 (0.0334)	0.143 (0.00882)
`condreg - LW`	0.00490 (0.0204)	0.00127 (0.0179)	-0.00204 (0.00446)
		$N_{\text{estim}} = 45$	
`sample`	-0.566 (0.105)	0.168 (0.0589)	0.230 (0.0146)
`LW`	-0.120 (0.0490)	0.149 (0.0366)	0.140 (0.00793)
`condreg`	-0.130 (0.0561)	0.151 (0.0314)	0.141 (0.00947)
`condreg - LW`	-0.0105 (0.0269)	0.00188 (0.0189)	0.000835 (0.00453)
		$N_{\text{estim}} = 60$	
`sample`	-0.275 (0.0592)	0.142 (0.0459)	0.175 (0.0100)
`LW`	-0.111 (0.0457)	0.147 (0.0357)	0.138 (0.00719)
`condreg`	-0.120 (0.0527)	0.154 (0.0310)	0.140 (0.00908)
`condreg - LW`	-0.120 (0.0527)	0.00696 (0.0158)	0.00161 (0.00386)

I.5 Turnover

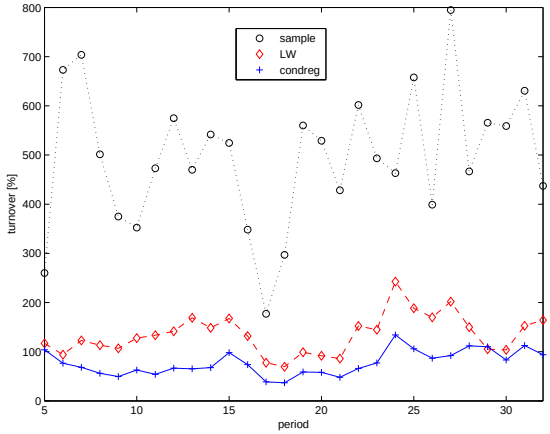
Turnover for various estimation horizon sizes over the trading period from February 18, 1994 through July 14, 2008. `sample`=sample covariance matrix, `LW`=linear shrinkage using the Ledoit-Wolf optimality, `condreg`=condition number regularization.



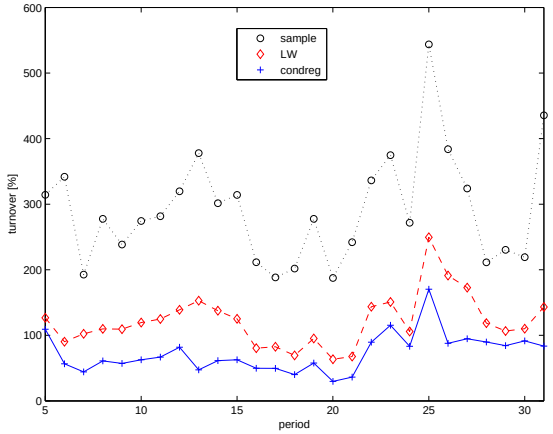
(a) $N_{\text{estim}} = 15$



(b) $N_{\text{estim}} = 30$



(c) $N_{\text{estim}} = 45$



(d) $N_{\text{estim}} = 60$

References

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

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

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.

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