Biometrika (2020), **107**, 2, *pp.* 397–414
Printed in Great Britain
doi: 10.1093/biomet/asz076
Advance Access publication 11 February 2020

Lassoing eigenvalues

BY DAVID E. TYLER

Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, New Jersey 08854, U.S.A.
dtyler@stat.rutgers.edu

AND MENGXI YI

School of Statistics, University of International Business and Economics, No. 10, Huixin Dongjie Beijing 100029, China
myi@uibe.edu.cn

SUMMARY

The properties of penalized sample covariance matrices depend on the choice of the penalty function. In this paper, we introduce a class of nonsmooth penalty functions for the sample covariance matrix and demonstrate how their use results in a grouping of the estimated eigenvalues. We refer to the proposed method as lassoing eigenvalues, or the elasso.

Some key words: Cross-validation; Geodesic convexity; Marchenko–Pastur distribution; Penalization; Principal component; Spiked covariance matrix.

1. INTRODUCTION AND MOTIVATION

Eigenvalues play a central role in many multivariate statistical methods. In working with the sample principal component roots, i.e., the eigenvalues of the sample covariance matrix, it has long been recognized that the larger roots tend to be overestimated and the smaller roots tend to be underestimated. Consequently, numerous methods have been proposed for shrinking eigenvalues together, including bias-correction (Anderson, 1965), decision-theoretic (Stein, 1975; Haff, 1991), Bayesian (Haff, 1980; Yang & Berger, 1994) and marginal-likelihood (Muirhead, 1982) approaches.

The aim of this paper is to study penalization methods for shrinking eigenvalues towards each other based on nonsmooth penalties. The rationale for using a nonsmooth penalty function is that the resulting penalization method not only can shrink the eigenvalues towards each other, but also can partition the eigenvalues into subgroups of equal values, i.e., the eigenvalues are lassoed together.

Partitioning the principal component roots into distinct groups can be viewed as a type of model selection method, with each of the 2^{q-1} possible partitions representing a different model. Here q denotes the dimension of the data. Models for which the p smallest eigenvalues, $p < q$, are taken to be equal are commonly referred to as sub-spherical models, factor models or reduced-rank covariance models (Anderson, 2003; Davis et al., 2014). A more general case in which subsets of eigenvalues can be equal is the well-studied spiked covariance model (Johnstone, 2001; Baik & Silverstein, 2006; Paul, 2007; Mestre, 2008; Bai & Yao, 2012). In general, taking subsets of the eigenvalues to be equal can yield covariance models with considerably fewer parameters than in the unrestricted covariance case. An obvious example is the case in which all the eigenvalues are

taken to be equal, which corresponds to the covariance being proportional to the identity matrix. In this case, the $q(q+1)/2$ distinct elements of a covariance matrix of order q are reduced to one parameter.

Rather than consider all possible 2^{q-1} partitions of the eigenvalues, our proposed penalized method reduces the set of models to q hierarchical models, with the first model being the case in which all eigenvalues are distinct, and the last model being the case in which all eigenvalues are equal. In the fixed-q setting, we show that the correct partition will almost surely be one of these q hierarchical models as the sample size n goes to infinity. In the setting where $q/n \to c$, we show that the model consistency property holds for spiked covariance models having r distinct roots, where r is fixed, and with the remaining $q - r$ roots being equal.

2. PENALIZED LIKELIHOOD ESTIMATES OF THE COVARIANCE MATRIX

2.1. *Preliminaries*

Let $X = \{x_1, \ldots, x_n\}$ be a q-dimensional sample of size n, with $\bar{x}$ representing its sample mean and $S_n = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^{\mathrm{T}}$ its sample covariance matrix. When S_n is nonsingular, which occurs with probability 1 for $n > q$ under random sampling from a continuous multivariate distribution, $(\bar{x}, S_n)$ uniquely minimizes

$$l(\mu, \Sigma; X) = n \log\{\det(\Sigma)\} + \sum_{i=1}^{n} (x_i - \mu)^{\mathrm{T}} \Sigma^{-1} (x_i - \mu) \tag{1}$$

over all $\mu \in \mathbb{R}^q$ and $\Sigma > 0$, i.e., the class of positive-definite symmetric matrices of order q. The function $l(\mu, \Sigma; X)$ corresponds, up to an additive constant, to two times the negative loglikelihood function under random sampling from a multivariate normal distribution. For singular S_n, which always occurs for $n \leqslant q$, the function $l(\mu, \Sigma; X)$ is not bounded from below.

Even when $n > q$, the sample covariance matrix is not very stable for small or even moderate values of n. Consequently, penalized sample covariance matrices have been introduced (Huang et al., 2006; Warton, 2008). Since penalizing the covariance matrix does not affect the estimate for μ, we consider the loss function

$$l(\Sigma; S_n) = n^{-1} l(\bar{x}, \Sigma; X) = \mathrm{tr}(\Sigma^{-1} S_n) + \log\{\det(\Sigma)\}, \tag{2}$$

which is uniquely minimized over $\Sigma > 0$ at S_n when $S_n > 0$. A penalized sample covariance matrix, say $\hat{\Sigma}_\eta$, is then defined as a minimizer over $\Sigma > 0$ of the penalized loss function

$$L(\Sigma; S_n, \eta) = l(\Sigma; S_n) + \eta \, \Pi(\Sigma). \tag{3}$$

Here $\Pi(\Sigma)$, defined on $\Sigma > 0$, denotes a nonnegative penalty function, and $\eta \geqslant 0$ is a tuning constant. Since $l(\Sigma; S_n)$ is strictly convex in Σ^{-1}, so is $L(\Sigma; S_n, \eta)$ when the penalty function is convex in Σ^{-1}. In this case the minimizer is uniquely defined when $S_n > 0$, with $\hat{\Sigma}_\eta$ being a continuous function of η.

When using the penalized approach, shrinking eigenvalues towards each other without penalizing the scale of the covariance matrix implies the use of a scale-invariant penalty, such that $\Pi(\Sigma) = \Pi(\gamma \Sigma)$ for $\Sigma > 0$ and $\gamma > 0$. The only scale-invariant penalty which is convex in Σ^{-1} is a constant penalty. For penalties that are not convex in Σ^{-1}, the uniqueness of a solution to (3) is not immediate, nor do convex optimization methods necessarily apply.

2.2. *Geodesically convex penalties*

A perhaps lesser known property of the negative loglikelihood function (2) is that it is strictly geodesically convex for any $S_n \neq 0$. This follows from a special case of Theorem 1 in Zhang et al. (2013). The concept of geodesic convexity, or g-convexity for short, is based on viewing the set of symmetric positive-definite matrices of order q as a Riemannian manifold with the geodesic path from $\Sigma_0 > 0$ to $\Sigma_1 > 0$ given by $\Sigma_t = \Sigma_0^{1/2}\{\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\}^t\Sigma_0^{1/2}$ for $0 \leqslant t \leqslant 1$; see Bhatia (2009) or Wiesel & Zhang (2015) for more details. A function $f(\Sigma)$ is then said to be g-convex if and only if $f(\Sigma_t) \leqslant (1-t)f(\Sigma_0) + tf(\Sigma_1)$ for $0 < t < 1$, and it is strictly g-convex if strict inequality holds for $\Sigma_0 \neq \Sigma_1$. Unlike convexity in Σ^{-1}, a function which is g-convex in Σ is also g-convex in Σ^{-1}.

Consequently, if one uses a g-convex penalty, then the penalized loss function $L(\Sigma; S_n, \eta)$ is strictly g-convex. For g-convex penalties, it has been shown that when $S_n > 0$, $L(\Sigma; S_n, \eta)$ has a unique minimizer $\hat{\Sigma}_\eta > 0$ which is a continuous function of $\eta \geqslant 0$. The above result also holds for singular S_n, provided some additional conditions are placed on the penalty function; for more details see Tyler & Yi (2019, Lemmas 2.2 and 2.3).

The Kullback–Liebler penalty $\Pi_{\mathrm{KL}}(\Sigma) = \mathrm{tr}(\Sigma^{-1}) + \log\{\det(\Sigma)\} - q$, corresponding to twice the Kullback–Liebler distance between Σ and I_q under the multivariate normal, is g-convex as well as convex in Σ^{-1}. A g-convex penalty which is not convex in Σ^{-1} is the Riemannian penalty $\Pi_{\mathrm{R}}(\Sigma) = \|\log\Sigma\|_{\mathrm{F}}^2$, which corresponds to the squared Riemannian distance between Σ and the identity matrix I_q. Here the norm $\|\cdot\|_{\mathrm{F}}$ refers to the Frobenius norm.

Although scale-invariant penalties cannot be convex in Σ^{-1}, they can be g-convex. In particular, given any g-convex penalty $\Pi(\Sigma)$, a scale-invariant penalty can be constructed by applying the penalty to the shape matrix $V(\Sigma) = \Sigma/\det(\Sigma)^{1/q}$. The new penalty $\Pi_{\mathrm{s}}(\Sigma) \equiv \Pi\{V(\Sigma)\}$ is scale invariant and also g-convex (Tyler & Yi, 2019, Lemma 4.1.i). The scale-invariant version of the Kullback–Liebler penalty is $\Pi_{\mathrm{s,KL}}(\Sigma) = \mathrm{tr}(\Sigma^{-1})\det(\Sigma)^{1/q} - q$, which is a measure of the distance between the geometric and harmonic means of the eigenvalues of Σ, while the Riemannian shape penalty is $\Pi_{\mathrm{s,R}}(\Sigma) = \|\log\Sigma - q^{-1}(\log\det\Sigma)\,I_q\|_{\mathrm{F}}^2$. Another example of a scale-invariant g-convex penalty is the condition number of Σ, i.e., the ratio of its largest eigenvalue to the smallest one.

2.3. *Orthogonally invariant penalties*

When considering penalties that have the effect of shrinking eigenvalues towards each other, it is natural to focus on those that are scale invariant and attain their minimum at any $\Sigma \propto I_q$, as well as penalties that depend on Σ only through its eigenvalues. This last property is equivalent to using an orthogonally invariant penalty function, i.e., one such that $\Pi(\Sigma) = \Pi(Q\Sigma Q^{\mathrm{T}})$ for any $Q \in \mathcal{O}(q)$, the group of orthogonal matrices of order q. Hereafter, the ordered eigenvalues of Σ will be denoted by $\lambda_1 \geqslant \cdots \geqslant \lambda_q > 0$ and the ordered eigenvalues of S_n by $d_1 \geqslant \cdots \geqslant d_q \geqslant 0$. Also, using the spectral value decomposition, one can express S_n as $P_n D_n P_n^{\mathrm{T}}$ with $P_n \in \mathcal{O}(q)$ and $D_n = \mathrm{diag}\{d_1, \ldots, d_q\}$.

In general, establishing that a function is g-convex can be challenging. However, for orthogonally invariant functions, it has recently been shown that g-convexity can be characterized as follows (Tyler & Yi, 2019, Theorem 3.1).

LEMMA 1. *The function $\Pi(\Sigma)$ is orthogonally invariant if and only if for some symmetric, i.e., permutation-invariant, function $\pi : \mathbb{R}^q \to \mathbb{R}$ one has $\Pi(\Sigma) = \pi\{\log(\lambda_1), \ldots, \log(\lambda_q)\}$, where $\lambda_1 \geqslant \cdots \geqslant \lambda_q > 0$ are the ordered eigenvalues of Σ. Furthermore, $\Pi(\Sigma)$ is (strictly) g-convex if and only if the function $\pi(y)$ is (strictly) convex.*

In addition, when using a g-convex orthogonally invariant penalty, the optimization problem (3) reduces to a convex optimization problem on the eigenvalues (Tyler & Yi, 2019; Theorem 5.1). Specifically, the minimum of $L(\Sigma; S_n, \eta)$ is attained at $\hat{\Sigma}_\eta = P_n \hat{\Lambda}_{n,\eta} P_n^{\mathrm{T}}$, where $\hat{\Lambda}_{n,\eta} = \mathrm{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_q\}$, with the diagonal terms corresponding to the minimizer over $\lambda_1 \geqslant \cdots \geqslant \lambda_q > 0$ of the function

$$\mathcal{L}(\lambda; d, \eta) = \sum_{j=1}^{q} \{d_j/\lambda_j + \log(\lambda_j)\} + \eta\, \pi\{\log(\lambda_1), \ldots, \log(\lambda_q)\}. \tag{4}$$

For $S_n \neq 0$ and a g-convex penalty, the function $\mathcal{L}(\lambda; d, \eta)$ is strictly convex in $y = \log \lambda \in \mathbb{R}^q$ for any $\eta \geqslant 0$. This follows from Lemma 1 since $\exp(-x)$ is strictly convex and $d_1 > 0$. Hence, if (4) admits a local minimum, then it corresponds to the unique global minimum. A minimum exists if and only if $\mathcal{L}(\lambda; d, \eta) \to \infty$ whenever $\|\log \lambda\| \to \infty$, which holds if $d_q > 0$, i.e., for $S_n > 0$. Furthermore, since eigenvalues are continuous functions of their matrix argument, it follows that the solution $\hat{\Lambda}_{n,\eta}$ is a continuous function of η.

3. NONSMOOTH PENALTY FUNCTIONS

The choice of the penalty term $\Pi(\Sigma)$ and tuning constant η determines the way in and extent to which the eigenvalues are shrunk towards each other. In this paper, we study the following class of nonsmooth penalty functions:

$$\Pi(\Sigma; a) = \sum_{j=1}^{q} a_j \log(\lambda_j), \quad a_1 \geqslant \cdots \geqslant a_q, \ \sum_{j=1}^{q} a_i = 0. \tag{5}$$

These not only shrink the roots together, but also generate equality for various subsets of eigenvalues for a large enough tuning constant. The penalties in (5), although continuous, are not differentiable in general since ordered eigenvalues are not differentiable functions at points of multiple roots.

The motivation for (5) came from first considering the special case of $\sum_{j<k} |\log(\lambda_j) - \log(\lambda_k)|$, which corresponds to choosing $a_1 = q-1, a_2 = q-3, \ldots, a_q = -(q-1)$. The absolute value signs in the penalty term are not necessary, of course, since $\lambda_j \geqslant \lambda_k$ for $j < k$, but are included to help relate the penalty to the l_1 penalty used in the regression lasso method. Other members of this class of penalty functions are discussed in § 4.

The function $\Pi(\Sigma; a)$ is scale invariant and orthogonally invariant as well as g-convex. The last property follows from applying the following lemma, proved in the Appendix, in conjunction with Lemma 1.

LEMMA 2. *For $a_1 \geqslant \cdots \geqslant a_q$, the function $\pi(y; a) = \sum_{j=1}^{q} a_j y_{(j)}$ is convex and symmetric, where $y_{(1)} \geqslant \cdots \geqslant y_{(q)}$ are the ordered values of $y \in \mathbb{R}^q$. Furthermore, if $\sum_{j=1}^{q} a_i = 0$, then $\pi(y; a) \geqslant 0$ with equality if and only if $y_1 = \cdots = y_q$.*

Observe that $\Pi(\Sigma; a) = \pi(\log \lambda; a)$. If we had simply defined $\pi(y; a) = \sum_{j=1}^{q} a_i y_j$, then although this is a convex function and in particular linear, it is not symmetric and so would not satisfy the conditions of Lemma 1. If the coefficients $a_1 \geqslant \cdots \geqslant a_q$ do not sum to zero, then $\Pi(\Sigma; a)$ would still be orthogonally invariant and g-convex, but would not be scale invariant.

Hereafter, unless stated otherwise, we focus on the case of $S_n > 0$. The case where S_n is singular is discussed in Remark 1. From the discussion in § 2.3, the problem of minimizing $L(\Sigma; S_n, \eta)$ over $\Sigma > 0$ when using the penalty $\Pi(\Sigma; a)$ reduces to the problem of minimizing

$$\mathcal{L}(\lambda; d, \eta) = \sum_{j=1}^{q} \{d_j/\lambda_j + (1 + \eta\, a_j) \log(\lambda_j)\} \tag{6}$$

subject to $\lambda_1 \geqslant \cdots \geqslant \lambda_q > 0$. To solve this optimization problem, first suppose that the solution satisfies $\hat{\lambda}_1 > \cdots > \hat{\lambda}_q > 0$, i.e., the minimum occurs at a point where all the eigenvalues are distinct and nonzero. In this case, owing to strict convexity, the solution corresponds to the unique critical point of (6), which is $\hat{\lambda}_j = d_j/(1 + \eta a_j)$ $(j = 1, \ldots q)$. If this solution does not satisfy $\hat{\lambda}_1 > \cdots > \hat{\lambda}_q > 0$, which will eventually be the case as η increases, then the true minimizer must contain at least one multiple root.

More generally, suppose that the minimum of (6) is achieved at a point where there are r different eigenvalues of Σ, say $\lambda_{(1)} > \cdots > \lambda_{(r)} > 0$ with respective multiplicities $m_1, \ldots, m_r$, so that $m_1 + \cdots + m_r = q$. Let $\mathcal{G} = \{G(1), \ldots, G(r)\}$ denote the corresponding partition of $\{1, \ldots, q\}$, i.e., $G(k) = (m_0 + \cdots + m_{k-1} + 1, \ldots, m_1 + \cdots + m_k)$ with $m_0 = 0$. Given the assumed multiplicities, the objective function (6) becomes

$$\mathcal{L}_{\mathcal{G}}(\lambda_{(1)}, \ldots \lambda_{(r)}; \tilde{d}, \eta) = \sum_{k=1}^{r} \{\tilde{d}_k/\lambda_{(k)} + (1 + \eta\, \tilde{a}_k) \log(\lambda_{(k)})\}, \tag{7}$$

where $\tilde{d}_k = \{\sum_{j \in G(k)} d_j\}/m_k$ and $\tilde{a}_k = \{\sum_{j \in G(k)} a_j\}/m_k$. If $\mathcal{G}$ is the correct partition, then (6) achieves its minimum at the unique critical point of $\mathcal{L}_{\mathcal{G}}$. This is given by

$$\hat{\lambda}_{(k)}(\mathcal{G}) = \tilde{d}_k/\{1 + \eta\, \tilde{a}_k\} \quad (k = 1, \ldots, r). \tag{8}$$

Conditions on η are needed, though, for this solution to satisfy the proper ordering.

LEMMA 3. *For $r > 1$, the solution (8) satisfies the constraint $\hat{\lambda}_{(1)}(\mathcal{G}) > \cdots > \hat{\lambda}_{(r)}(\mathcal{G}) > 0$ if and only if $\eta < \eta(\mathcal{G}) = \inf\{\tilde{\eta}_k(\mathcal{G}) : k = 1, \ldots r - 1\}$, where*

$$\tilde{\eta}_k(\mathcal{G}) = \frac{\tilde{d}_k - \tilde{d}_{k+1}}{\tilde{a}_k \tilde{d}_{k+1} - \tilde{a}_{k+1} \tilde{d}_k}$$

if $\tilde{a}_k \tilde{d}_{k+1} > \tilde{a}_{k+1} \tilde{d}_k$ and $\tilde{\eta}_k = \infty$ otherwise. For $r = 1$, the solution $\hat{\lambda}_{(1)}(\mathcal{G}) = \bar{d}$ is valid for any $\eta < \infty$.

The condition $\eta < \eta(\mathcal{G})$ is a necessary but not sufficient condition for $\mathcal{G}$ to be the correct partition. It is possible for more than one partition to satisfy $\eta < \eta(\mathcal{G})$; in particular, this condition is always satisfied when $r = 1$. It remains, then, to find the correct partition $\mathcal{G}$. For a given η, the minimizer of $\mathcal{L}(\lambda; d, \eta)$ must correspond to the minimizer of $\mathcal{L}_{\mathcal{G}}(\lambda_{(1)}, \ldots, \lambda_{(r)}; \tilde{d}, \eta)$ for some $\mathcal{G}$ such that $\eta < \eta(\mathcal{G})$.

It is not necessary to check all 2^{q-1} partitions of $\{1, \ldots, q\}$ to find the unique minimizer of $\mathcal{L}(\lambda; d, \eta)$. Rather, the unique minimizer can be found by considering only the following q hierarchical partitions. Let $\mathcal{G}_q = \{\{1\}, \ldots, \{q\}\}$. For $\eta < \eta(\mathcal{G}_q)$, it readily follows that $\mathcal{G}_q$ is the minimizing partition. Next, define $\mathcal{G}_{q-1}$ to be the partition formed by joining the two eigenvalues

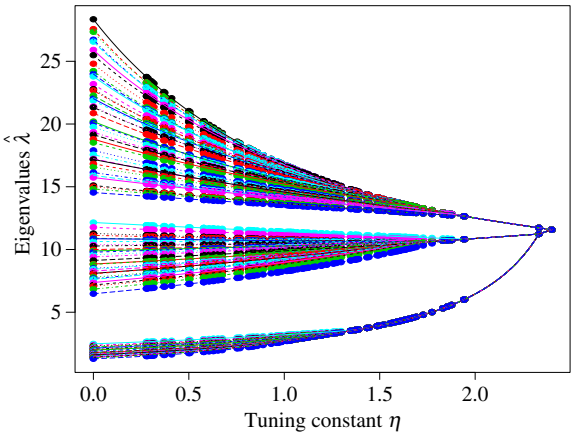

Fig. 1. An example of the elasso using Marchenko–Pastur weights.

that become equal at $\eta = \eta(\mathcal{G}_q)$. Continue in this fashion to produce the sequence of partitions $\mathcal{G}_q, \ldots, \mathcal{G}_1$, with $\mathcal{G}_1 = \{\{1, \ldots, q\}\}$. Specifically, given $\mathcal{G}_r = \{G_r(1), \ldots, G_r(r)\}$, define

$$\mathcal{G}_{r-1} = \{G_r(1), \ldots, G_r(k_r^* - 1),\ G_r(k_r^*) \cup G_r(k_r^* + 1),\ G_r(k_r^* + 2), \ldots, G_r(r)\}, \tag{9}$$

where $k_r^* = \arg\inf\{k \mid \tilde{\eta}_k(\mathcal{G}_r),\ k = 1, \ldots, r-1\}$. Using this notation, we characterize the minimizer of $\mathcal{L}(\lambda; d, \eta)$ in the following theorem.

THEOREM 1. *Suppose* $d_1 > \cdots > d_q > 0$ *and that* k_r^* *defined in* (9) *is unique for each* $r = 2, \ldots, q$. *Then* $0 < \eta(\mathcal{G}_q) < \cdots < \eta(\mathcal{G}_1) \equiv \infty$. *Furthermore, for* $\eta(\mathcal{G}_{r+1}) \leqslant \eta < \eta(\mathcal{G}_r)$ *with* $\eta(\mathcal{G}_{q+1}) \equiv 0$, $\mathcal{L}(\lambda; d, \eta) \geqslant \mathcal{L}(\hat{\lambda}; d, \eta)$ *where* $\hat{\lambda}_j = \hat{\lambda}_{(k)}(\mathcal{G}_r)$ *for* $j \in G_r(k)$. *Consequently, the unique minimizer of* $L(\Sigma; S_n; \eta)$ *when* $\Pi(\Sigma) = \Pi(\Sigma; a)$ *is* $\hat{\Sigma}_\eta = P_n \hat{\Lambda}_{n,\eta} P_n^{\mathrm{T}}$ *where* $\hat{\Lambda}_{n,\eta} = \mathrm{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_q\}$. *Moreover,* $\hat{\Sigma}_\eta$ *is continuous in* η.

Remark 1. The conditions in Theorem 1 hold with probability 1 when sampling from a continuous distribution. The conditions that $d_1 > \cdots > d_q > 0$ and that k_r^* be unique are not necessary, but are included so that the values of $\eta(\mathcal{G}_r)$ will all be distinct. An extension of this theorem, which includes the population version of the elasso and the $n \leqslant q$ case, is given in the Appendix. When S_n is singular, we have existence and uniqueness of $\hat{\Sigma}_\eta$ for large enough η.

When using $\Pi(\Sigma; a)$, we refer to the penalized method of estimating the covariance matrix as the elasso. The elasso has a number of properties similar to the lasso for regression. The estimated precision matrix $\hat{\Sigma}_\eta^{-1}$ is a piecewise-linear function of η, with the q knots or kinks in the function occurring at $0 = \eta_{(q)} < \cdots < \eta_{(1)} < \infty$, where $\eta_{(r)} = \eta(\mathcal{G}_{r+1})$. Hence, only the knots and the values of $\hat{\lambda}_1, \ldots, \hat{\lambda}_q$ at the knots, together with P_n, need to be known to reconstruct the value of $\hat{\Sigma}_\eta$ for all values of η. The locations of the knots $\eta_{(k)}$ are easy to compute, and unlike with the regression lasso, the value of $\hat{\lambda}$ at a knot has a simple closed form; in particular it is a linear function of the sample eigenvalues. The knots of the elasso yield a hierarchical set of q models, $\mathcal{G}_q \succ \cdots \succ \mathcal{G}_1$, where $\mathcal{G}_a \succ \mathcal{G}_b$ means that the sets in $\mathcal{G}_b$ can be formed by taking unions of sets in $\mathcal{G}_a$. In general, for $\eta_{(r)} \leqslant \eta < \eta_{(r-1)}$, where $\eta_{(0)} \equiv \infty$, the grouping of the eigenvalues of $\hat{\Sigma}_\eta$ consists of the r groups indicated by the partition $\mathcal{G}_r$.

To illustrate the elasso, a pedagogical example is given in Fig. 1, which shows a simulated sample of size $n = 1000$ from a 100-dimensional multivariate normal distribution, for which the

covariance matrix has 40 eigenvalues equal to 20, 30 eigenvalues equal to 10, and 30 eigenvalues equal to 2. The choice of the weights $a_1, \ldots, a_q$ used in the example is based on the Marchenko–Pastur law. These weights are discussed in § 5; see (11). The points displayed in Fig. 1 correspond to the knots at which two eigenvalue groups are joined. Any eigenvalues that are joined at a given knot remain joined for all η greater than that knot, thus producing the eigenvalue tree and paths seen in the figure. The eigenvalue tree gives 100 possible models or groupings of the eigenvalues. In this simulation, the correct grouping of the roots, i.e., with the multiplicities 40, 30 and 30, occurs at the third-from-last knot.

4. Choice of weights and path consistency

When using the elasso, values for the weights $a_1, \ldots, a_q$ and the tuning constant η need to be chosen. The choice of weights depends partly on the application of interest. Consider the log-condition-number penalty $\log(\lambda_1/\lambda_q)$, which corresponds to $a_1 = 1$, $a_2 = \cdots = a_{q-1} = 0$ and $a_q = -1$. This penalty lassoes only a group of the largest eigenvalues together and/or a group of the smallest eigenvalues together for any fixed η, as illustrated in Fig. 2(a). The condition number has been considered by other authors in the context of constrained likelihood problems (e.g., Wiesel, 2012; Won et al., 2013) but has not previously been studied as a penalty term.

To obtain more general groupings of the roots, the weights a_j should all be different. As η increases, the solutions behave in a manner similar to that displayed in Fig. 1, i.e., two groups of roots come together at each knot until all the roots become equal. Also, as stated in the following theorem, the elasso path is then strongly consistent for fixed q, i.e., the probability of the path eventually containing the correct model tends to 1 as $n \to \infty$. The proof is given in the Supplementary Material.

Theorem 2. *Let $x_1, \ldots, x_n$ be a random sample from x, a q-dimensional distribution with mean μ_o and covariance matrix Σ_o, where the multiplicities of the eigenvalues of Σ_o correspond to the partition $\mathcal{G}_o$. For the penalty $\Pi(\Sigma; a)$ defined by (5), if $a_1 > \cdots > a_q$, then* $\mathrm{pr}(\mathcal{G}_o \in \{\mathcal{G}_q, \ldots, \mathcal{G}_1\}$ *for large enough n*$) = 1$, *where $\mathcal{G}_j$ ($j = q, \ldots, 1$) is as defined in (9).*

Hence, of the 2^{q-1} possible models for eigenvalue multiplicities, for large n there is a high probability that the correct model is one of the q models in the path.

The extreme sample roots are known to be more heavily biased than the less extreme roots. Consequently, the penalty originally used to motivate the elasso, $\sum_{j<k} |\log(\lambda_j) - \log(\lambda_k)|$, tends not to sufficiently penalize the extreme roots for modest sample sizes, as can be seen in Fig. 2(b). A more promising choice of weights can be motivated as follows. The two roots joined at the first knot in the elasso are d_{j^*} and d_{j^*+1}, where j^* corresponds to the index for which the value of $\kappa_j = (d_j - d_{j+1})/(a_j d_{j+1} - a_{j+1} d_j)$ is minimized, but not negative, over $j = 1, \ldots, q-1$. When $\Sigma \propto I_q$, it would be desirable for the values of κ_j to be nearly equal. This would hold if $a_j \approx \hat{a}_j = (d_j - \bar{d})/\bar{d}$, since then $\kappa_j \approx 1$ for $j = 1, \ldots, q$. Furthermore, the knot at which all roots are made equal in the elasso is

$$\eta_{(1)} = \sup\{\eta_k^* : k = 1, \ldots q-1\}, \quad \eta_k^* = \frac{\bar{d}_k - \bar{d}}{\bar{a}_k - \bar{d}}, \tag{10}$$

with $\bar{d}_k = \sum_{j=1}^k d_j/k$ and $\bar{a}_k = \sum_{j=1}^k a_j/k$. The partition into two groups before all roots are made equal in the elasso is given by $\mathcal{G}_2 = \{(1, \ldots, k_o), (k_o + 1, \ldots, q)\}$, where k_o is the value of

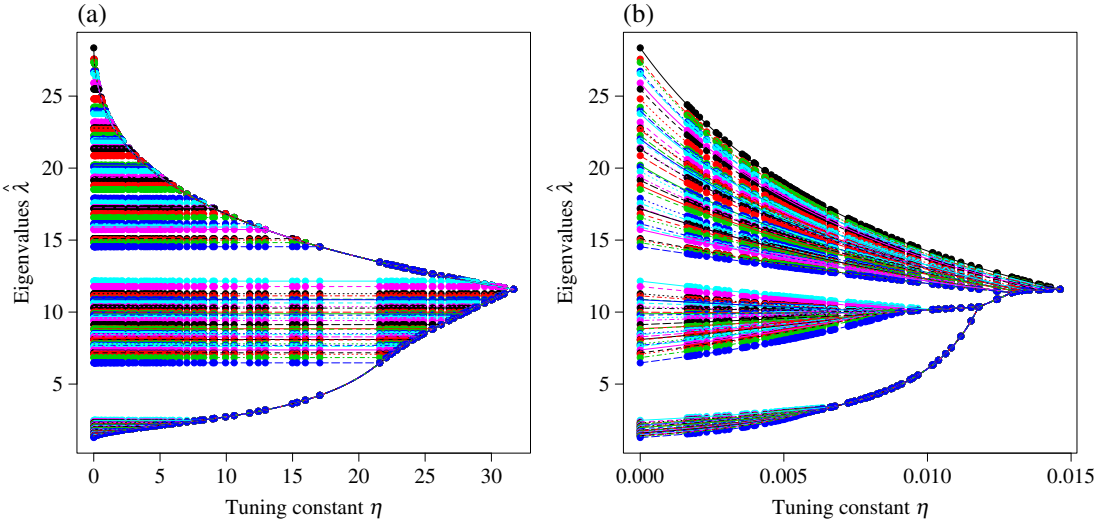

Fig. 2. (a) An example of the elasso using the log-condition-number weights; (b) the penalty $\sum_{j<k} |\log(\lambda_j) - \log(\lambda_k)|$ for the data used in Fig. 1.

k that maximizes η_k^* in (10). When $\Sigma \propto I_q$, it would also be desirable for the values of η_k^* to be nearly equal. This again occurs when $a_j \approx \hat{a}_j$, which gives $\eta_k^* \approx 1$ for $k = 1, \ldots, q-1$.

To obtain weights which satisfy $a_j \approx \hat{a}_j$ when $\Sigma \propto I_q$, one could use the mean or the median of the distribution of $\hat{a}_j$, say under spherical normality. However, this distribution is not particularly tractable, though it can be simulated for given q and n. As shown in the next section, for the large-q, large-n setting, the Marchenko & Pastur (1967) law can be used to generate weights that approximate $\hat{a}_j$ whenever $\Sigma \propto I_q$.

5. MARCHENKO–PASTUR WEIGHTS

The Marchenko–Pastur law arises in the following way. Suppose that $x_1, \ldots, x_n$ is a random sample of $x \in \mathbb{R}^q$, where x itself has q identical and independent components with unit variance and finite fourth moments. Let F_q denote the empirical distribution of the eigenvalues $d_1 \geqslant \cdots \geqslant d_q$ of the sample covariance matrix S_n, i.e., $F_q(x) = \#\{d_i \leqslant x\}/q$. Under the setting $n \to \infty$ and $q/n \to \nu$, $F_q(x) \to F_{\mathrm{MP}}(x; \nu)$ almost surely, where $F_{\mathrm{MP}}(x; \nu)$ is the Marchenko–Pastur distribution function with parameter ν. For $\nu \in (0, 1]$, the density of the Marchenko–Pastur distribution is $f_{\mathrm{MP}}(x; \nu) = (2\pi x \nu)^{-1}\{(c_+ - x)(x - c_-)\}^{1/2}$ with support $c_- \leqslant x \leqslant c_+$, where $c_\pm = (1 \pm \sqrt{\nu})^2$.

We define the Marchenko–Pastur weights used in Fig. 1 by centring the decreasing quantiles of the Marchenko–Pastur law:

$$a_{\mathrm{MP},j} = \xi_j - \bar{\xi}, \qquad \xi_j = F_{\mathrm{MP}}^{-1}\{(q - j + 0.5)/q;\, q/n\}. \tag{11}$$

For $\hat{a}_j = (d_j - \bar{d})/\bar{d}$ and under the asymptotic setting of the Marchenko–Pastur law,

$$\max_{j=1,\ldots,q} \left|\hat{a}_j - a_{\mathrm{MP},j}\right| \to 0 \tag{12}$$

almost surely. The above statements remain true if the unit variance in the Marchenko–Pastur law is replaced by a constant variance σ^2. As conjectured earlier, this result allows us to show that all

the roots are made equal in the elasso at approximately $\eta = 1$ when using the Marchenko–Pastur weights. Consequently, if the tuning parameter is chosen so that $\eta > 1$, then one obtains the estimate $\hat{\Sigma}_\eta = \bar{d}\, I_q$ almost surely whenever the true model is $\Sigma \propto I_q$.

THEOREM 3. *For $n > q$, suppose that $x_1, \ldots, x_n$ is a random sample from $x \in \mathbb{R}^q$, where x itself has q identical and independent components with variance σ^2 and finite fourth moments. Consider the knots from the elasso based on the Marchenko–Pastur weights $a_{\mathrm{MP},j}$. For any fixed $k \geqslant 2$, as $n \to \infty$ and $q/n \to \nu \in (0, 1)$, the last knot $\eta_{(1)} \to 1$ almost surely.*

Asymptotic results on the behaviour of the elasso using the Marchenko–Pastur weights when the covariance matrix is not proportional to the identity can also be obtained. An extension of the Marchenko–Pastur law to a spiked covariance model, as stated in Baik & Silverstein (2006, Theorem 1.1), is the following. Suppose that $x_1, \ldots, x_n$ is a random sample from $x = Az$, with A nonsingular and $z \in \mathbb{R}^q$, having q identical and independent components with unit variance and finite fourth moments. Also, suppose that the eigenvalues of the covariance matrix AA^{T} of x are

$$\lambda_1 \geqslant \cdots \geqslant \lambda_r > \lambda_{r+1} = \cdots = \lambda_{q-s} = \sigma^2 > \lambda_{q-s+1} \geqslant \cdots \geqslant \lambda_q, \tag{13}$$

where $\lambda_r/\sigma^2 > 1 + \sqrt{\nu}$ and $\lambda_{q-s+1}/\sigma^2 < 1 - \sqrt{\nu}$. As $n \to \infty$ and $q/n \to \nu \in (0, 1)$, with r and s held fixed, the Marchenko–Pastur law applies to the distribution of the standardized base roots, i.e., to d_j/σ^2 for $j \in \mathcal{S}_{\mathrm{b}} = \{r + 1, \ldots, q - s\}$. Also, the spikes

$$d_j \to \lambda_j^{\mathrm{o}} \equiv \lambda_j\{1 + \nu\sigma^2/(\lambda_j - \sigma^2)\} \tag{14}$$

almost surely for $j \notin \mathcal{S}_{\mathrm{b}}$. These results allow us to establish the following generalization of Theorem 3.

THEOREM 4. *For $n > q$, suppose that $x_1, \ldots, x_n$ is a random sample from $x = Az \in \mathbb{R}^q$, where z itself has q identical and independent components with unit variance and finite fourth moments. Assume that the eigenvalues of AA^{T} are given by* (13).

(i) *If all the nonbase roots have multiplicity* 1, *i.e., $\lambda_1 > \cdots > \lambda_r$ and $\lambda_{q-s+1} > \cdots > \lambda_q$, then for any fixed $k \geqslant 2$, the knot $\eta_{(r+s+1)} \to 1$ almost surely as $n \to \infty$ with $q/n \to \nu \in (0, 1)$. Furthermore, $\mathcal{G}_{r+s+2} = \{\{1\}, \ldots, \{r\}, \{r + 1, \ldots, q - s\}, \{q - s + 1\}, \ldots, \{q\}\}$.*

(ii) *In general, suppose that the eigenvalues* (13) *consist of $t + 1$ distinct roots with various multiplicities, and let $\mathcal{G}_{\mathrm{o}}$ denote the partition of $\{1, \ldots, q\}$ into the $t + 1$ subsets associated with these multiplicities. Then there exists an integer m with $t \leqslant m \leqslant r + s + 1$ such that $\eta_{(m)} \to 1$ almost surely as $n \to \infty$ with $q/n \to \nu \in (0, 1)$. Furthermore, $\{r + 1, \ldots, q - s\} \in \mathcal{G}_{(m+1)}$, with all other elements of $\mathcal{G}_{m+1}$ being subsets of elements in $\mathcal{G}_{\mathrm{o}}$.*

Theorem 4 says that if the spiked roots are all unique, then one of the q hierarchical models in the elasso path, specifically $\mathcal{G}_{r+s+2}$ which forms at $\eta_{(r+s+1)} = \eta(\mathcal{G}_{r+s+2})$, is almost surely the true model. When the spiked roots have multiplicities greater than 1, then the true model may not be in the elasso path. Nevertheless, as stated in (ii), there almost surely exists a model in the elasso path that separates the base roots from the spike roots, with two or more spike roots being grouped together only if they correspond to the same spike. This is illustrated in Fig. 3(b), while Fig. 3(a) illustrates Theorem 3.

Under the spiked covariance model, when using the elasso with the Marchenko–Pastur weights, if one chooses the tuning parameter to be $\eta = 1 + \epsilon$ for small enough $\epsilon > 0$, then the roots associated with the base space are almost surely identified. The value of $\epsilon > 0$ needed to

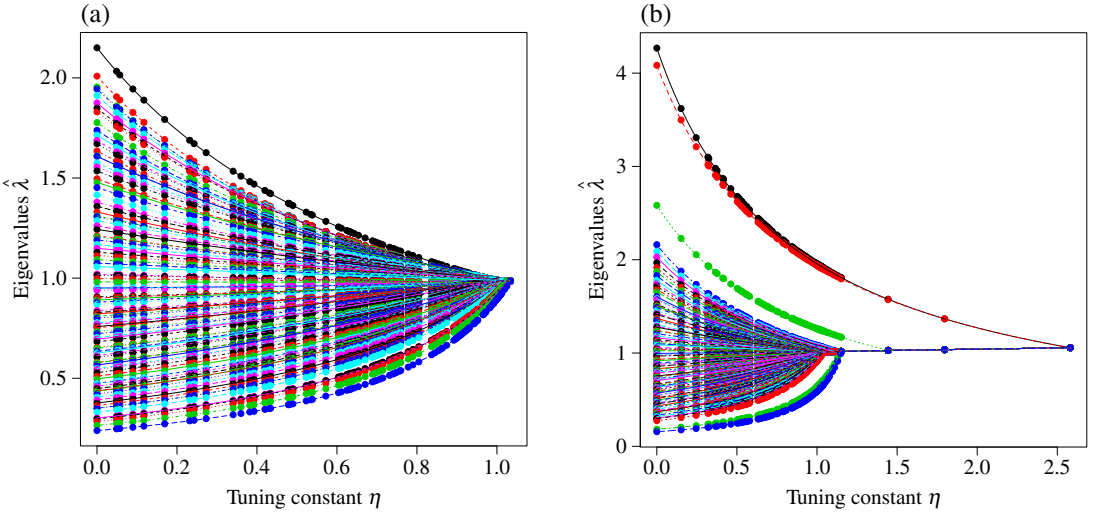

Fig. 3. Examples of elassos using weights a_{MP} for multivariate normal samples with $q = 100$ and $n = 400$; the eigenvalues of the covariance matrix are (a) $\lambda_1 = \cdots = \lambda_q = 1$ and (b) $\lambda_1 = \lambda_2 = 4$, $\lambda_3 = 2$, $\lambda_4 = \cdots = \lambda_{q-2} = 1$ and $\lambda_{q-1} = \lambda_q = 0.25$.

obtain these results depends on how much separation there is between the base root σ^2 and the spikes. It is known that some separation between the base root and the spikes is needed in order to distinguish the spike roots from the base root. In particular, the condition stated after (13), namely $\lambda_r/\sigma^2 > 1 + \sqrt{\nu}$ and $\lambda_{q-s+1}/\sigma^2 < 1 - \sqrt{\nu}$, is a necessary condition (Baik & Silverstein, 2006). The condition on the spikes needed to almost surely distinguish the spike roots from the base root at $\eta = 1 + \epsilon$ is

$$\lambda_r/\sigma^2 > 1 + \sqrt{\nu} + o(\epsilon), \quad \lambda_{q-s+1}/\sigma^2 < 1 - \sqrt{\nu} - o(\epsilon). \tag{15}$$

The proofs of (15), (12), and Theorems 3 and 4 are given in the Supplementary Material.

Table 1 reports the results from a simulation study of the distribution of the last knot of the elasso under spherical normal sampling when using the Marchenko–Pastur weights. The convergence of the value of the last knot to 1 follows from the discussion above. Also, the asymptotic value as $n \to \infty$ with q fixed tends to provide a better approximation to the last knot at given q and n than the asymptotic value under the setting $n \to \infty$ with $q/n \to \nu$. To obtain the estimate $\hat{\Sigma}_\eta = \bar{d}\, I_q$ with high probability, whenever spherical normality holds, one could choose η to be, say, three standard deviations above the mean of the distribution of the last knot. For $q = 100$ and $n = 500$, this gives $\eta \approx 1.10$. By Theorem 4, this would also result in a good chance of separating the base root from the spikes in the spiked covariance model. However, such a choice may not be appropriate outside the spiked covariance model, as in the example of Fig. 1. Consequently, in the next section we consider tuning the elasso via the data-driven method of cross-validation.

6. CROSS-VALIDATION AND MODEL SELECTION

For penalized approaches, cross-validation can be applied to the unpenalized objective function, i.e., (1) in this setting, which gives

$$\mathrm{CV}(\eta; \mathcal{A}) = n_{\mathcal{A}} \log\{\det(\hat{\Sigma}_{-\mathcal{A},\eta})\} + \sum_{x_i \in \mathcal{A}} (x_i - \bar{x}_{-\mathcal{A}})^{\mathrm{T}}\, \hat{\Sigma}^{-1}_{-\mathcal{A},\eta} (x_i - \bar{x}_{-\mathcal{A}}),$$

Table 1. *Simulation under spherical normality for the last knot of the elasso when using Marchenko–Pastur weights; the mean and standard deviation (in parentheses) of the distribution of the last knot, based on* 1000 *simulations, are given over q and* $\mathrm{df} = n - 1$

$q \setminus \mathrm{df}$	q	$2q$	$5q$	$10q$	$50q$	$100q$	$1000q$
10	1.141	1.142	1.151	1.149	1.155	1.148	1.142
	(0.135)	(0.133)	(0.132)	(0.127)	(0.125)	(0.125)	(0.124)
30	1.082	1.076	1.076	1.074	1.076	1.074	1.074
	(0.073)	(0.067)	(0.059)	(0.056)	(0.053)	(0.051)	(0.0497)
50	1.057	1.055	1.056	1.057	1.056	1.054	1.054
	(0.051)	(0.044)	(0.041)	(0.038)	(0.036)	(0.036)	(0.037)
100	1.038	1.036	1.034	1.034	1.033	1.035	1.034
	(0.033)	(0.031)	(0.025)	(0.023)	(0.022)	(0.022)	(0.022)
300	1.019	1.018	1.017	1.017	1.016	1.017	1.016
	(0.017)	(0.014)	(0.012)	(0.012)	(0.010)	(0.011)	(0.010)
500	1.013	1.012	1.012	1.012	1.012	1.012	1.012
	(0.012)	(0.010)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)
1000	1.008	1.008	1.008	1.008	1.008	1.008	1.008
	(0.007)	(0.006)	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)

calculated for a range of η values (Stone, 1974; Huang et al., 2006). Here $\mathcal{A}$ denotes a subset of the data, with $\bar{x}_{-\mathcal{A}}$ and $\hat{\Sigma}_{-\mathcal{A},\eta}$ representing, respectively, the sample mean vector and the penalized estimate of Σ not based on the data in $\mathcal{A}$. K-fold cross-validation then seeks to minimize $K^{-1}\sum_{k=1}^{K} \mathrm{CV}(\eta; \mathcal{A}_k)$ over $\eta \geqslant 0$, where $\mathcal{A}_1, \ldots, \mathcal{A}_K$ is a random partition of the data into subsets of equal size, plus or minus one.

Figure 4(a) shows the results of five-fold cross-validation for the data and weights used in Fig. 1. The black curve in the middle represents the mean of the five values of $\mathrm{CV}(\eta; \mathcal{A})$, and the blue curves correspond to $\pm$ one standard error of the mean of these five values. One hundred evenly spaced values between 0 and 2.5 are used for η. The minimum value in the plot is 63 005, which is obtained at $\eta = 0.675$. Given the partition used in the simulations, namely three distinct roots with multiplicities 40, 30 and 30, it can be observed that the grouping of the eigenvalues in Fig. 1 at $\eta = 0.675$ is too coarse.

In regression lasso, a relaxed lasso is often recommended (Meinshausen, 2007) in order to obtain a simpler model. The analogy for the elasso would be to choose a larger value of η having a cross-validation mean equal to the cross-validation plus one standard error at $\eta = 0.675$, which in this case corresponds to $\eta = 1.075$. Again, this does not yield a refined enough partition. The reason why cross-validation does not do well at selecting the correct partition of the roots is that the correct partition does not arise until $\eta = 1.95$. At this point, although the partition is correct, the roots are overly shrunk together and so the estimates of the eigenvalues result in a poor fit.

A proposed modification is demonstrated in Fig. 4(b). For each of the 100 partitions or models in the original elasso path, five-fold cross-validation is applied to an elasso under the corresponding model. The elasso for a given partition $\mathcal{G}_r = \{G(1), \ldots, G(r)\}$, as defined in (9), is obtained by minimizing (7) over $\lambda_{(1)} > \cdots > \lambda_{(r)}$; details are given in the Supplementary Material. The graph plots the minimum value of the cross-validation against the corresponding model knot. Here, the smallest model cross-validation error occurs at the correct partitioning of the eigenvalues, i.e., at the eigenvalue multiplicities of 40, 30 and 30.

When using an elasso for a given partition in the original elasso path, the resulting path is identical to the original path once the given partition is reached. Hence, no partitions are obtained

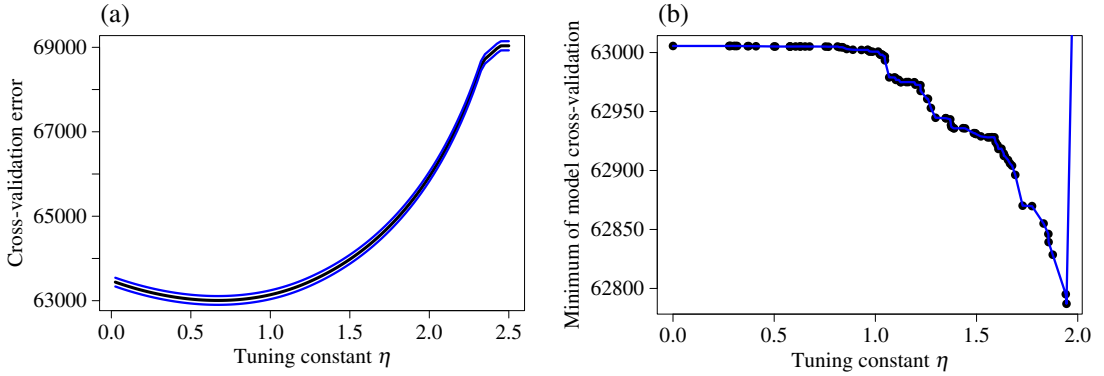

Fig. 4. Results of (a) five-fold cross-validation and (b) model cross-validation for the simulated data used in Fig. 1; each knot corresponds to a different model.

which are not in the original path. Since the second stage is not concerned with model selection, penalties other than an elasso penalty can be used in this stage. However, the results of the simulation study reported in the next section do not seem to be heavily dependent on the choice of estimator used in the second stage.

7. SIMULATIONS, A DATA EXAMPLE AND DISCUSSION

7.1. *A simulation study*

The following estimators of the covariance matrix are compared in a simulation study: S, the sample covariance matrix; LW, the estimator of Ledoit & Wolf (2004); F, an estimator proposed by Friedman (1989); $\mathrm{E_{CV}}$, the elasso estimator with η determined by cross-validation; $\mathrm{E_{MCV}}$, the elasso estimator obtained via model cross-validation; $\mathrm{E_R}$, the elasso estimator with the Riemannian shape penalty used at the model stage; and $\mathrm{E_F}$, the elasso estimator with Friedman's estimator used at the model stage.

The Marchenko–Pastur weights are used in all of the elassos, and all cross-validations are five-fold. Simulation results for other choices of weights in the elasso are reported in the Supplementary Material. Both the Ledoit–Wolf and the Friedman covariance estimators are of the form $(1-\beta)S_n + \beta\bar{d}I_q$. They differ in that for the estimator proposed by Friedman (1989) the value of β is chosen by cross-validation, while for the estimator of Ledoit & Wolf (2004) it is determined using a consistent estimate for its optimal value; see Ledoit & Wolf (2004) for details. Friedman's estimator for a partition $\mathcal{G}$ is defined to be $(1-\beta)\hat{\Sigma}_{\mathcal{G}}+\beta\bar{d}I_q$, where $\hat{\Sigma}_{\mathcal{G}}$ is the maximum likelihood estimator under the partition $\mathcal{G}$, i.e., $\hat{\Sigma}_{\mathcal{G}}$ is obtained by replacing the roots in S_n with the average of the sample roots of their corresponding set in $\mathcal{G}$. The Riemannian shape penalty was discussed at the end of § 2.2.

Samples of size $n = 100$ from a multivariate normal distribution of dimension $q = 30$ were simulated for the five different covariance models given below. To evaluate the estimators at a given model, in addition to the two criteria used in Huang et al. (2006), namely the Kullback–Liebler or entropy loss $\mathrm{KL} = \mathrm{tr}(\Sigma^{-1}\hat{\Sigma}) - \log\det(\Sigma^{-1}\hat{\Sigma}) - q$ and the quadratic loss $D^2 = \mathrm{tr}\{(\Sigma^{-1}\hat{\Sigma} - I_q)^2\}$, we also consider the Riemannian loss $R = \|\log(\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2})\|_{\mathrm{F}}$. Since all the estimators considered are scale and orthogonally equivariant, and all the loss functions depend only on the eigenvalues of $\Sigma^{-1}\hat{\Sigma}$, it follows that the results of the simulation still hold if any particular Σ is replaced by $\lambda P\Sigma P^{\mathrm{T}}$ for $\lambda > 0$ and $P \in \mathcal{O}(q)$.

Table 2. *Simulation results for Models* 1–5*; reported are the means, with* * *indicating the minimum, and standard deviations (in parentheses) calculated over* 100 *simulation runs*

Loss	Model	S	LW	F	E_{CV}	E_{MCV}	E_R	E_F
KL	Model 1	5.314	0.027	0.019	0.010*	0.020	0.060	0.063
		(0.316)	(0.032)	(0.021)	(0.013)	(0.027)	(0.055)	(0.052)
	Model 2	5.314	34.709	35.212	2.237	0.736	0.675	0.625*
		(0.316)	(2.434)	(2.877)	(0.659)	(0.267)	(0.191)	(0.227)
	Model 3	5.314	4.254	3.376	0.441	0.317*	0.325	0.326
		(0.316)	(0.268)	(0.268)	(0.100)	(0.085)	(0.102)	(0.101)
	Model 4	5.314	10.493	4.690	4.215	3.087*	3.660	3.741
		(0.316)	(0.734)	(0.366)	(0.373)	(0.296)	(0.296)	(0.344)
	Model 5	5.314	4.128	4.214	3.936	3.438*	3.964	4.014
		(0.316)	(0.298)	(0.360)	(0.407)	(0.325)	(0.373)	(0.330)
D^2	Model 1	9.369	0.054	0.039	0.020*	0.042	0.122	0.127
		(0.763)	(0.066)	(0.042)	(0.025)	(0.061)	(0.117)	(0.114)
	Model 2	9.369	1471.508	1514.158	14.706	2.825	1.610*	2.250
		(0.763)	(196.518)	(235.344)	(6.772)	(1.455)	(0.580)	(1.138)
	Model 3	9.369	8.938	10.366	0.514*	0.703	0.747	0.746
		(0.763)	(0.785)	(1.125)	(0.094)	(0.222)	(0.321)	(0.316)
	Model 4	9.369	39.767	12.983	9.050	5.687*	6.948	9.254
		(0.763)	(4.166)	(1.420)	(1.106)	(0.649)	(0.741)	(1.069)
	Model 5	9.369	11.584	11.954	8.114	6.642*	10.399	10.982
		(0.763)	(1.177)	(1.420)	(1.235)	(0.905)	(1.392)	(1.303)
R	Model 1	3.555	0.193	0.168	0.117*	0.165	0.305	0.323
		(0.133)	(0.129)	(0.103)	(0.083)	(0.115)	(0.167)	(0.147)
	Model 2	3.555	3.698	3.709	1.569	1.037	1.131	0.972*
		(0.133)	(0.047)	(0.057)	(0.168)	(0.153)	(0.151)	(0.141)
	Model 3	3.555	3.017	2.358	1.078	0.789*	0.794	0.794
		(0.133)	(0.122)	(0.077)	(0.145)	(0.110)	(0.114)	(0.113)
	Model 4	3.555	3.974	2.920	2.954	2.613*	2.856	2.684
		(0.133)	(0.099)	(0.108)	(0.129)	(0.138)	(0.136)	(0.130)
	Model 5	3.555	2.713	2.731	2.897	2.743	2.727	2.696*
		(0.133)	(0.090)	(0.103)	(0.135)	(0.125)	(0.114)	(0.100)

S, the sample covariance matrix; LW, the estimator of Ledoit & Wolf (2004); F, an estimator proposed by Friedman (1989); ECV, the elasso estimator with $\eta_\cdot$ determined by cross-validation; EMCV, the elasso estimator obtained via model cross-validation; ER, the elasso estimator with the Riemannian shape penalty used at the model stage; and EF, the elasso estimator with Friedman's estimator used at the model stage.

Model 1: $\Sigma = I_q$.
Model 2: $\Sigma^{-1} = \{\gamma_{ij}\}$, where $\gamma_{ii} = 1$ and $\gamma_{ij} = 0.6$ if $i \neq j$.
This model has eigenvalues $\lambda_1 = \cdots = \lambda_{29} = 2.25$ and $\lambda_{30} = 0.05435$.
Model 3: $\Sigma = \sigma^2\{(1-\rho)I + \rho 1_q 1_q^{\mathrm{T}}\}$ with $\sigma^2 = 0.5$ and $\rho = 0.7$.
This model has eigenvalues $\lambda_1 = 10.65$ and $\lambda_2 = \cdots = \lambda_{30} = 0.15$.
Model 4: $\Sigma = \mathrm{diag}\{\lambda_1, \cdots, \lambda_{30}\}$, with $\lambda_1 = \cdots = \lambda_5 = 20$, $\lambda_6 = \cdots = \lambda_{15} = 10$ and $\lambda_{16} = \cdots = \lambda_{30} = 1$.
Model 5: $\Sigma = \{\sigma_{ij}\}$ with $\sigma_{ii} = 1$ for $i = 1, \ldots, q$ and $\sigma_{ij} = \rho^{|i-j|}$ for $i \neq j$, where $\rho = 0.7$.
This covariance matrix arises from an AR(1) model and its eigenvalues are all distinct.

The simulations were repeated over 100 runs, and the means and standard deviations of the simulated losses are reported in Table 2. In general, the elasso E_{CV} and the two-stage elassos

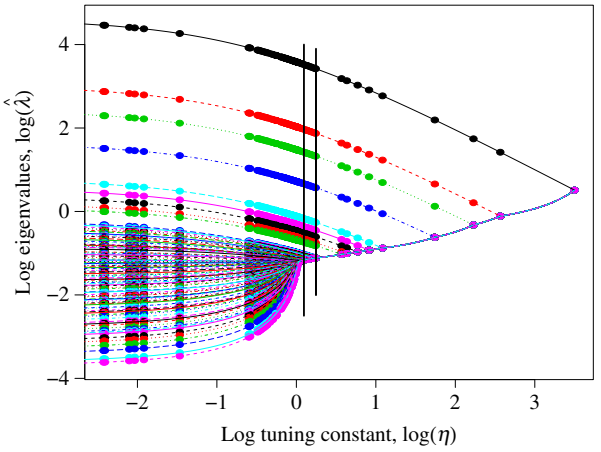

Fig. 5. The elasso results, plotted on a log-log scale, for the call centre data using the Marchenko–Pastur weights. The first vertical line represents the values of η obtained via cross-validation, and the second vertical line corresponds to the model obtained via model cross-validation.

E_{MCV}, E_R and E_F perform better than the other estimators, with the difference being particularly pronounced for Models 2, 3 and 4. The two-stage elasso estimators tend to perform slightly worse than the elasso estimator for Model 1. We suspect this is due to overfitting by the two-stage elasso estimators. Despite this overfitting, the two-stage elassos tend to result in better estimators for the other models.

7.2. *Telephone call centre data*

As an application example, we consider the call centre data previously analysed by Huang et al. (2006), among others. On each weekday in 2002, except for holidays and six days during which the data-collection equipment was out of operation, telephone calls were recorded from 7:00 a.m. until midnight, yielding a sample of size 239. For each of these days, the responses correspond to the number of calls received in consecutive 10-minute periods, resulting in a 102-dimensional response vector N. Since the number of calls tends not to be normally distributed, each data point is transformed by $x = (N + 0.25)^{1/2}$, where the operation acts on each of the elements of x and N. The sample $x_1, \ldots, x_{239}$ is presumed to be a set of independent observations.

Huang et al. (2006) calculated penalized covariance estimates using a penalty defined on a modified Cholesky decomposition of the covariance matrix and tuned via five-fold cross-validation. Their calculations are based on a training set consisting of the first $n = 205$ data points; therefore, to make our analysis comparable, we also consider only the first $n = 205$ data points and use five-fold cross-validation. Figure 5 shows the results obtained from the elasso when using the Marchenko–Pastur weights.

For the Marchenko–Pastur weights, the minimum five-fold cross-validation mean is 490.2, with a standard error of 242.3, which is attained at $\eta = 1.1$, i.e., $\log(\eta) = 0.095$. For the elasso, the model at $\eta = 1.1$ corresponds to a spiked covariance model with the 19 largest eigenvalues having multiplicity 1 and the smallest eigenvalue having multiplicity 83. The minimum of the five-fold model cross-validation is 371.7, and the corresponding model is a spiked covariance model with the nine largest eigenvalues having multiplicity 1 and the smallest eigenvalue having multiplicity 93.

By comparison, the cross-validation mean is 5436.3 for the sample covariance matrix and 1822.0 for the Ledoit–Wolf estimator. For estimators tuned via cross-validation, the minimum cross-validation mean is 3168.3 for the penalized estimator proposed by Huang et al. (2006), 1457.8 for the elasso estimator based on the log-condition-number penalty, 608.1 for the Riemannian shape penalized estimator, 519.2 for Freidman's estimator, and 436.05 for the elasso estimator based on the penalty $\sum_{j<k} |\log \lambda_j - \log \lambda_k|$.

7.3. *Further discussion*

The results of our simulations and analysis of a real dataset suggest that if one uses an elasso penalty with cross-validation, the resulting elasso covariance estimator can yield significant improvements in performance over other covariance estimators. Based on theoretical arguments given in § 5, as well as the simulation study presented in the Supplementary Material, we recommend choosing the Marchenko–Pastur weights for the elasso penalty. Furthermore, we recommend using the two-stage model elasso estimators.

Analogous to the lasso in regression, an important feature of our proposed method is that the elasso path gives a set of q hierarchical models for the multiplicities of the eigenvalues of the covariance matrix. This can be helpful in gaining a basic understanding of the structure of the covariance matrix, without explicitly assuming a parsimonious model for Σ. As previously noted, the model $\mathcal{G}_1$, which corresponds to $\Sigma = \sigma^2 I_q$, contains only one parameter, as opposed to the $q(q+1)/2$ parameters in an unrestricted Σ. In general, the number of parameters for the covariance model associated with a partitioning of the eigenvalues into $g \leqslant q$ groups can be shown to be $q(q+1)/2 - m(m-1)/2 - (q-g)$, where $m \leqslant q-g+1$ represents the cardinality of the largest group. In the high-dimensional scenario of $m/q \to \tau$ as $q \to \infty$, the proportional reduction in parameters converges to $\tau^2 \times 100\%$, which is 100% for a spiked covariance model. Once a model for the multiplicities of the eigenvalues of the covariance matrix is obtained, one can focus on the eigenspaces associated with the groups of eigenvalues rather than on individual eigenvectors.

Acknowledgement

We thank Jianhua Huang and Haipeng Shen for providing us with the edited version of the call centre data analysed in this paper. The original dataset was made available to them by Avi Mandelbaum. An R (R Development Core Team, 2020) package to implement the elasso is currently being developed in collaboration with Klaus Nordhausen. Yi was supported in part by the Scientific Research Starting Foundation of the University of International Business and Economics, Beijing. Both authors were supported in part by the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of Theorems 2–4.

Appendix

Proofs

Proof of Lemma 2. The functions $\pi_r(y) = \sum_{j=1}^{r} y_{(j)}$ are convex for $r = 1, \ldots, q$, with $\pi_q(y) = \sum_{j=1}^{q} y_j$ being linear. The function $\pi(y; a)$ is symmetric and can be expressed as $\pi(y; a) = \sum_{r=1}^{q} b_r \pi_r(y)$ where

$b_r = a_r - a_{r+1} \geqslant 0$ for $r = 1, \ldots, q-1$ and $b_q = a_q$. Each of the summands $b_r \pi_r(y)$ is convex, and hence $\pi(y; a)$ is convex. When $\sum_{r=1}^{q} a_r = 0$, the function $\pi(y; a)$ is invariant under the transformation $y_j \to y_j + c$ $(j = 1, \ldots, q)$. So to show that $\pi(y; a)$ is nonnegative in this case, assume without loss of generality that $y_{(q)} = 0$. This implies $\pi_r(y) \geqslant 0$ and so $\pi(y; a) \geqslant \{\sum_{r=1}^{q-1} r b_r\} y_{(q)} = 0$, with equality if and only if $y_1 = \cdots = y_q$. □

Proof of Lemma 3. The inequality $\hat{\lambda}_k(\mathcal{G}) > \hat{\lambda}_{k+1}(\mathcal{G})$ holds if and only if $\tilde{d}_k - \tilde{d}_{k+1} > \eta(\tilde{a}_k \tilde{d}_{k+1} - \tilde{a}_{k+1} \tilde{d}_k)$, which holds if and only if $\eta < \hat{\eta}_k$. Hence, the inequality holds for all $k = 1, \ldots, r-1$ if and only if $\eta < \eta(\mathcal{G})$. □

Proof of Theorem 1. By definition (9), $\hat{\lambda}_k(\mathcal{G}_{r-1}) = \hat{\lambda}_k(\mathcal{G}_r)$ for $k < k_r^*$ and $\hat{\lambda}_k(\mathcal{G}_{r-1}) = \hat{\lambda}_{k+1}(\mathcal{G}_r)$ for $k > k_r^*$. Also, it can be shown that $\hat{\lambda}_{k_r^*}(\mathcal{G}_{r-1}) = \gamma \hat{\lambda}_{k_r^*}(\mathcal{G}_r) + (1-\gamma)\hat{\lambda}_{k_r^*+1}(\mathcal{G}_r)$ for some $0 < \gamma < 1$. Specifically, $\gamma = m_{k_r^*}(1 + \eta \tilde{a}_{k_r^*})/\{m_{k_r^*}(1 + \eta \tilde{a}_{k_r^*}) + m_{k_r^*+1}(1 + \eta \tilde{a}_{k_r^*+1})\}$, where $\tilde{a}_k$ and m_k are defined with respect to the partition $\mathcal{G}_r$. This implies that if $\hat{\lambda}_1(\mathcal{G}_r) > \cdots > \hat{\lambda}_r(\mathcal{G}_r)$, then $\hat{\lambda}_1(\mathcal{G}_{r-1}) > \cdots > \hat{\lambda}_{r-1}(\mathcal{G}_{r-1})$. By Lemma 3, the former holds if and only if $\eta < \eta(\mathcal{G}_r)$, and the latter holds if and only if $\eta < \eta(\mathcal{G}_{r-1})$. Thus $\eta(\mathcal{G}_r) < \eta(\mathcal{G}_{r-1})$, where the inequality is strict since it is assumed that k_r^* is well-defined.

If $0 \leqslant \eta < \eta(\mathcal{G}_q)$, it readily follows that $\hat{\lambda}_j = \hat{\lambda}_j(\mathcal{G}_q)$. We use finite induction to complete the proof. Suppose that for $\eta(\mathcal{G}_{r+2}) \leqslant \eta < \eta(\mathcal{G}_{r+1})$ we have $\hat{\lambda}_j = \hat{\lambda}_k(\mathcal{G}_{r+1})$ for $j \in G_{r+1}(k)$. It then follows from the continuity of the solution in η that for $\eta = \eta(\mathcal{G}_{r+1})$, the solution corresponds to $\hat{\lambda}_j = \hat{\lambda}_k(\mathcal{G}_r)$ for $j \in G_r(k)$. This solution also holds for any $\eta(\mathcal{G}_{r+1}) \leqslant \eta < \eta(\mathcal{G}_r)$, because otherwise, if $\mathcal{G}_r$ were not the optimizing partition for some η in the interval, there would be a discontinuity of the solution at that value of η. □

Extensions of Theorem 1

Although the conditions in Theorem 1 requiring that the eigenvalues of S_n be distinct and that k_r^* be unique hold with probability 1 under random sampling from a continuous multivariate distribution with $n > q$, they are not necessary. For the penalty $\Pi(\Sigma; a)$, consider the general problem of minimizing $L(\Sigma; \tilde{S}, \eta)$ over $\Sigma > 0$, where $\tilde{S} > 0$ is a given matrix which may not satisfy the above conditions. Theorem 1 then requires a slight modification, namely $0 \leqslant \eta(\mathcal{G}_q) \leqslant \cdots \leqslant \eta(\mathcal{G}_2) < \eta(\mathcal{G}_1) = \infty$; that is, the knots of the elasso are not necessarily unique. With this modification, the statement in Theorem 1 holds.

If the eigenvalues of $\tilde{S}$ form $p < q$ distinct groups, then $0 = \eta(\mathcal{G}_q) = \cdots = \eta(\mathcal{G}_{p+1}) < \eta(\mathcal{G}_p)$. For example, if $\tilde{S} \propto I$, then $0 = \eta(\mathcal{G}_q) = \cdots = \eta(\mathcal{G}_2) < \eta(\mathcal{G}_1) = \infty$. In general, if k_r^* is not unique, but rather the infimum in its definition, given after (9), is obtained at $t \leqslant r-1$ points, then t knots occur at the same point, i.e., $\eta(\mathcal{G}_r) = \cdots = \eta(\mathcal{G}_{r-t+1})$.

The above results can be used to define a model elasso associated with a given partition of the eigenvalues, say $\mathcal{G}_o$ with corresponding multiplicities $m_1, \ldots, m_p$ such that $m_1 + \cdots + m_p = q$. In other words, consider minimizing $L(\Sigma; S_n, \eta)$ over all $\Sigma > 0$ with the given multiplicities of the ordered eigenvalues. The solution to this minimization problem is the same as the solution to the problem of minimizing $L(\Sigma; \tilde{S}, \eta)$ over $\Sigma > 0$, where $\tilde{S}$ is the maximum likelihood estimate of Σ under $\mathcal{G}_o$. Here, $\tilde{S} = P_n \tilde{D} P_n^{\mathrm{T}}$ where $\tilde{D}$ is a diagonal matrix with elements $\tilde{d}_k$ repeated m_k times, for $k = 1, \ldots, p$. The resulting solution is then given by $\hat{\Sigma}_\eta = P_n \tilde{\Delta}_\eta P_n^{\mathrm{T}}$ where $\tilde{\Delta}_\eta$ is a diagonal matrix with elements corresponding to the solution to (7) when $r = p$. The multiplicities of the elements $\tilde{\Delta}_\eta$ do not necessarily correspond to the multiplicities of the elements of $\tilde{D}$ for large enough η, since different groups of roots are eventually joined together as η increases.

When the values of k_r^*, for $r \leqslant p$, are unique, there are p distinct knots of the model elasso, namely $0 = \eta_{(q)} = \cdots = \eta_{(p)} < \eta_{(p-1)} < \cdots < \eta_{(1)} < \infty$. If $\mathcal{G}_o$ corresponds to the one of the partitions generated by an unrestricted elasso, i.e., from $\mathcal{G}_q \succ \cdots \succ \mathcal{G}_1$, then the partitions generated by the restricted or model elasso are $\mathcal{G}_p \succ \cdots \succ \mathcal{G}_1$; that is, the partitions $\mathcal{G}_p, \ldots, \mathcal{G}_1$ are the same for the unrestricted and restricted elassos.

The elasso when S_n is singular

Suppose that S_n has rank $r < q$; then the form of $\hat{\Sigma}_\eta$ given in Theorem 1 still corresponds to the unique minimizer of $L(\Sigma; S_n, \eta)$ whenever $\eta > -1/\alpha_{r+1}$, where $\alpha_k = \sum_{j=k}^{q} a_j/(q-k+1) < 0$. For $0 \leqslant \eta < -1/\alpha_{r+1}$, though, a minimizer of $L(\Sigma; S_n, \eta)$ does not exist.

To justify these statements, recall that the elements of $\hat{\Lambda}_{n,\eta} = \mathrm{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_q\}$ correspond to the minimum of (9) over $\lambda_1 \geqslant \cdots \geqslant \lambda_q > 0$, which in the non-full-rank case is equivalent to minimizing

$$\mathcal{L}(\lambda; d, \eta) = \sum_{j=1}^{r} \{d_j/\lambda_j + (1+\eta\, a_j)\log(\lambda_j)\} + \sum_{j=r+1}^{q} (1+\eta\, a_j)\log(\lambda_j). \tag{A1}$$

The second sum can be rewritten as

$$(q-r)(1+\eta\,\alpha_{r+1})\log(\lambda_{r+1}) + \sum_{j=r+1}^{q} (1+\eta\, a_j)\log(\rho_j),$$

where $\rho_j = \lambda_j/\lambda_{r+1} \leqslant 1$ for $j = r+1, \ldots, q$. For $\eta < -1/\alpha_{r+1}$ we have $(1+\eta\,\alpha_{r+1}) > 0$, and so (A1) $\to -\infty$ as $\lambda_{r+1} \to 0$ with λ_j for $j = 1, \ldots, r$ and ρ_j for $j = r+1, \ldots, q$ held fixed. Hence $\mathcal{L}(\Sigma; S_n, \eta)$ does not have a minimum over $\Sigma > 0$ when $\eta < -1/\alpha_{r+1}$.

Now consider the case where $\eta > -1/\alpha_{r+1}$. To show that a minimizer of $L(\Sigma; S_n, \eta)$ over $\Sigma > 0$ exists, it is sufficient to show that (A1) is coercive, i.e., (A1) $\to \infty$ as $\lambda_q \to 0$ and/or $\lambda_1 \to \infty$. Consider the alternative expression for the second summand in (A1),

$$(q-r)(1+\eta\,\alpha_{r+1})\log(\lambda_r) + \sum_{j=r+1}^{q} (q-j+1)(1+\eta\,\alpha_j)\log(\beta_j), \tag{A2}$$

where $\beta_j = \lambda_j/\lambda_{j-1} \leqslant 1$, $j = 2, \ldots, q$. Since $\eta > -1/\alpha_{r+1}$ and $\alpha_k \geqslant \alpha_{k+1}$, it follows that $(1+\eta\,\alpha_j) < 0$, $j = r+1, \ldots, q$. If $\lambda_q \to 0$, then either $\beta_j \to 0$ for some $j = r+1, \ldots, q$ or $\lambda_r \to 0$. This implies (A2) $\to \infty$ as $\lambda_q \to 0$ provided λ_r is bounded above. Also, the first sum in (A1) is bounded below since it can be expressed as

$$\sum_{j=1}^{r} \{d_j/\lambda_j + (1+\eta\,\bar{a}_r)\log(\lambda_j)\} + \sum_{j=1}^{r} (a_j - \bar{a}_r)\log(\lambda_j),$$

with $\bar{a}_r = \sum_{j=1}^{r} a_j/r \geqslant 0$, where each term in the first sum is bounded from below and, by Lemma 2, the second sum is nonnegative. Therefore, (A1) $\to \infty$ as $\lambda_q \to 0$ provided λ_r is bounded above. If $\lambda_q \to 0$ and $\lambda_r \to \infty$, then we also have (A1) $\to \infty$ since

$$\sum_{j=1}^{r} (1+\eta\,\bar{a}_r)\log(\lambda_j) + (q-r)(1+\eta\,\alpha_{r+1})\log(\lambda_r) \geqslant q\log(\lambda_r) \to \infty.$$

It remains to consider the case $\lambda_1 \to \infty$ with λ_q being bounded away from 0. In this case, since $\sum_{j=1}^{r} d_j/\lambda_j \geqslant 0$, it is sufficient to show $\sum_{j=1}^{q}(1+\eta\, a_j)\log(\lambda_j) = \sum_{j=1}^{q}\log(\lambda_j) + \eta\,\pi(\log\lambda; a) \to \infty$, which readily follows since $\sum_{j=1}^{q}\log(\lambda_j) \to \infty$ and $\pi(\log\lambda; a) \geqslant 0$.

References

Anderson, G. A. (1965). An asymptotic expansion for the distribution of the latent roots of the estimated covariance matrix. *Ann. Math. Statist.* **36**, 1153–73.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis.* New York: Wiley.

BAI, Z. & YAO, J. (2012). On sample eigenvalues in a generalized spiked population model. *J. Mult. Anal.* **106**, 167–77.

BAIK, J. & SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Mult. Anal.* **97**, 1382–408.

BHATIA, R. (2009). *Positive Definite Matrices*. Princeton, New Jersey: Princeton University Press.

DAVIS, R. A., ZANG, P. & ZHENG, T. (2014). Reduced-rank covariance estimation in vector autoregressive modeling. *arXiv:* 1412.2183.

FRIEDMAN, J. H. (1989). Regularized discriminant analysis. *J. Am. Statist. Assoc.* **84**, 165–75.

HAFF, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8**, 586–97.

HAFF, L. R. (1991). The variational form of certain Bayes estimators. *Ann. Statist.* **19**, 1163–90.

HUANG, J. Z., LIU, N., POURAHMADI, M. & LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93**, 85–98.

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29**, 295–327.

LEDOIT, O. & WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Mult. Anal.* **88**, 365–411.

MARCHENKO, V. A. & PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.* **1**, 457–83.

MEINSHAUSEN, N. (2007). Relaxed lasso. *Comp. Statist. Data Anal.* **52**, 374–93.

MESTRE, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Trans. Info. Theory* **54**, 5113–29.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley.

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance matrix. *Statist. Sinica* **17**, 1617–42.

R DEVELOPMENT CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

STEIN, C. (1975). Estimation of a covariance matrix. Rietz Lecture, 39th Annual Meeting IMS, Atlanta, Georgia.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.* B **36**, 111–47.

TYLER, D. E. & YI, M. (2019). Shrinking the covariance matrix using convex penalties on the matrix-log transformation. *arXiv:* 1903.08281.

WARTON, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Am. Statist. Assoc.* **103**, 340–9.

WIESEL, A. (2012). Unified framework to regularized covariance estimation in scaled Gaussian models. *IEEE Trans. Sig. Proces.* **60**, 29–38.

WIESEL, A. & ZHANG, T. (2015). Structured robust covariance estimation. *Foundat. Trends Sig. Proces.* **8**, 127–216.

WON, J., LIM, J., KIM, S. & RAJARATNAM, B. (2013). Condition number regularized covariance estimation. *J. R. Statist. Soc.* B **75**, 427–50.

YANG, R. & BERGER, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22**, 1195–211.

ZHANG, T., WIESEL, A. & GRECO, M. S. (2013). Multivariate generalized Gaussian distribution: Convexity and graphical models. *IEEE Trans. Sig. Proces.* **61**, 4141–8.

[*Received on* 3 *May* 2018. *Editorial decision on* 17 *July* 2019]