

On sparsity scales and covariance matrix transformations

BY H. S. BATTEY

*Department of Mathematics, Imperial College London, South Kensington Campus,
 London SW7 2AZ, U.K.*

h.battey@imperial.ac.uk

SUMMARY

We develop a theory of covariance and concentration matrix estimation on any given or estimated sparsity scale when the matrix dimension is larger than the sample size. Nonstandard sparsity scales are justified when such matrices are nuisance parameters, distinct from interest parameters, which should always have a direct subject-matter interpretation. The matrix logarithmic and inverse scales are studied as special cases, with the corollary that a constrained optimization-based approach is unnecessary for estimating a sparse concentration matrix. It is shown through simulations that for large unstructured covariance matrices, there can be appreciable advantages to estimating a sparse approximation to the log-transformed covariance matrix and converting the conclusions back to the scale of interest.

Some key words: Concentration matrix; High-dimensional nuisance parameter; Matrix exponential; Matrix logarithm; Matrix transformation; Sparsity.

1. INTRODUCTION

The covariance matrix Σ^* and concentration matrix $\Omega^* = (\Sigma^*)^{-1}$ of a p -dimensional random vector X play a central role in summarizing linear relationships between variables and appear throughout multivariate analysis (e.g., Muirhead, 1982). When p is small relative to the sample size n , the sample covariance matrix S_n and its inverse are useful estimators of Σ^* and Ω^* , and their statistical properties are well understood. However, some modern applications yield data with $p \gg n$. Since the performance of statistical procedures is often quantified by their behaviour as n grows large, allowing $p \gg n$ entails a double asymptotic regime in which p grows at least as fast as n . While S_n remains elementwise consistent as p and n grow, S_n^{-1} does not exist, and noise accumulation leads to inconsistency in the stronger matrix norms relevant for applications.

For any $p \times p$ matrices A and B with eigenvalues $\lambda_k(A)$ and $\lambda_k(B)$ ($k = 1, \dots, p$), Weyl's inequality implies that

$$\max_k |\lambda_k(A) - \lambda_k(B)| \leq \|A - B\|_{\text{op}},$$

where $\|A\|_{\text{op}} = \sup_{\|x\|_2 \leq 1} \|Ax\|_2$ is the spectral norm. For eigenvectors $\xi_k(A)$ and $\xi_k(B)$, an application of the so-called $\sin \theta$ theorem of Davis & Kahan (1970) gives

$$\|\xi_k(A) - \xi_k(B)\|_2 \leq \frac{\sqrt{2}\|A - B\|_{\text{op}}}{\min\{|\lambda_{k-1}(A) - \lambda_k(B)|, |\lambda_k(B) - \lambda_{k+1}(A)|\}}.$$

Thus, under conditions on the eigenvalue spacings, convergence in the spectral norm is sufficient for convergence of eigenvectors and eigenvalues. Implications of spectral norm consistency

$(n, p \rightarrow \infty)$ for, inter alia, discriminant analysis and linear regression are given in § 3 and in the Supplementary Material.

By contrast, elementwise consistency is too weak for many applications and does not imply convergence of eigenvectors and eigenvalues, a fact exemplified by the sample covariance matrix, which is elementwise consistent as n and p tend to infinity provided that $(\log p)/n \rightarrow 0$, but has severely overdispersed eigenvalues relative to its population counterpart if p grows proportionally to n (Marchenko & Pastur, 1967).

The assumption necessary for progress is sparsity, and consistent $(n, p \rightarrow \infty)$ estimators of an elementwise-sparse Σ^* and Ω^* have been proposed by Bickel & Levina (2008) and Cai et al. (2016). But consistency is only interesting insofar as the assumptions made are satisfied to an adequate order of approximation. The present paper is motivated by the observation that in many statistical contexts, for example in linear regression and discriminant analysis, a covariance or concentration matrix is a nuisance parameter. Interpretability of Σ^* or Ω^* is then immaterial and less important than consistency of the interpretable object of interest. It seems sensible to model explicitly only those aspects of direct concern and retain a level of agnosticism over the nuisance parameters, which suggests a data-dependent sparsity scale for estimation of Σ^* or Ω^* .

Let $\mathcal{V}_p^+$ be the space of symmetric positive-definite $p \times p$ matrices and $\mathcal{V}_p$ a general space of $p \times p$ matrices. In the vein of Box & Cox (1964), it is natural to seek a data-dependent bijective transformation $f : \mathcal{V}_p^+ \rightarrow \mathcal{V}_p$ of Σ^* such that $f(\Sigma^*)$ is of maximal sparsity, to make a more efficient and valid analysis on the transformed scale by exploiting this sparsity, and to convert the conclusions back to the scale of interest.

A reasonable way to define a parametric family of matrix transformations of Σ^* is through its eigenvalues, with the matrix of orthonormal eigenvectors held fixed. For instance, $\log(\Sigma^*) = \Xi^* \log(\Lambda^*) \Xi^{*T}$, where $\log(\Lambda^*)$ is the diagonal matrix of log-transformed eigenvalues of Σ^* and Ξ^* is the matrix of corresponding orthonormal eigenvectors. Thus consider, for instance, the two-parameter family of eigenvalue transformations

$$\lambda^{(a,b)} = (\log \lambda)^b (1 + ab \log \lambda)^{1/b}, \quad a \in [-1, 1], \quad b \in [0, 1], \quad (1)$$

with special cases $\lim_{b \rightarrow 0} \lambda^{(1,b)} = \lambda$, $\lim_{b \rightarrow 0} \lambda^{(-1,b)} = \lambda^{-1}$ and $\lambda^{(0,1)} = \log \lambda$. These generate the matrix transformations $\Sigma^* \mapsto \Sigma^*$, $\Sigma^* \mapsto (\Sigma^*)^{-1}$ and $\Sigma^* \mapsto \log(\Sigma^*)$, where the latter coincides (e.g., Battey, 2017, proof of Lemma 1.1) with the definition of $L^* = \log(\Sigma^*)$ through a Taylor series expansion of the matrix exponential, $\Sigma^* = \exp(L^*) = \sum_{k=0}^{\infty} (L^*)^k / k!$.

Ideally, one would estimate the sparsity scale, and, in principle, this is feasible using (1) or similar and a likelihood-based criterion. Framing this as a convex optimization problem seems challenging, but in the present paper we assume that a suitable, ideally optimal, sparsity scale has been determined and establish the estimation theory on that basis, with focus on the matrix logarithm in numerical examples.

The following notation is used. The eigenvalues are ordered from largest to smallest so that, for example, the maximum and minimum eigenvalues of a matrix A are $\lambda_1(A)$ and $\lambda_p(A)$. We simplify this notation for specific widely used matrices. For instance, the eigenvalues of Σ^* are written $\lambda_1^*, \dots, \lambda_p^*$. The Frobenius norm and elementwise norm of A are written $\|A\|_F^2$ and $\|A\|_{\max}$, and are defined as $\text{trace}(A^T A) = \sum_{u,v} |a_{uv}|^2$ and $\max_{u,v} |a_{uv}|$. If the matrix A is symmetric, the spectral norm simplifies to $\|A\|_{\text{op}} = |\lambda_1(A)|$ and

$$\|A\|_{\text{op}} \leq \max_u \sum_v |a_{uv}|.$$

For a vector, $x = (x_1, \dots, x_p)^T$, the ℓ_0 -norm of x is $\|x\|_0 = \sum_{v=1}^p \mathbb{I}(x_j \neq 0)$, where $\mathbb{I}(\mathcal{E})$ is the indicator of the event $\mathcal{E}$, taking value 1 if $\mathcal{E}$ is true and 0 otherwise. The notation $A \succ 0$ and $A \succeq 0$ mean, respectively, that A is strictly positive definite and positive semidefinite. The canonical basis vectors for $\mathbb{R}^p$ are written $e_1, \dots, e_p$. For a positive integer p , the set $\{1, \dots, p\}$ is written $[p]$. Finally, $\asymp$ means equality in order of magnitude, and $a \vee b$ and $a \wedge b$ denote the maximum and minimum of scalars a and b .

Proofs are given in the Supplementary Material, except when their inclusion in the main article brings appreciable insight. To avoid unnecessary technical details, we assume that Σ^* is the covariance matrix of a $p \times 1$ random vector X of zero mean.

2. ESTIMATION UNDER SPARSITY ON A TRANSFORMED SCALE

2.1. General discussion

Assume that the function $f : \mathcal{V}_p^+ \rightarrow \mathcal{V}_p$ of § 1 is known or has been estimated. Then $f(\Sigma^*) = \Xi^* f(\Lambda^*) \Xi^{*T}$ where $\Lambda^* = \text{diag}(\lambda_1^*, \dots, \lambda_p^*)$ is the diagonal matrix of ordered eigenvalues of Σ^* and $\Xi^* = (\xi_1^*, \dots, \xi_p^*)$ is the matrix of corresponding orthonormal eigenvectors. The matrix-valued function f is thus defined by the same scalar-valued function applied to the eigenvalues elementwise and yields symmetric matrices by construction.

Write $F^* = f(\Sigma^*)$. A simple estimator of a sparse F^* is the thresholding estimator $\tilde{F}$, which entails setting all entries of a pilot estimator $\hat{F}^P$ of F^* to zero if they are below a predefined threshold ν in absolute value. This procedure depends for its success on a skilful choice of pilot estimator. In particular, $\hat{F}^P$ must converge elementwise to F^* , as we show in § 2.3. Thus there are two main challenges associated with estimation under sparsity on a transformed scale: given a positive-definite pilot estimator $\hat{\Sigma}^+$ of Σ^* , under what conditions and at what rate will $\hat{F}^P = f(\hat{\Sigma}^+)$ converge to F^* elementwise? Given a spectral norm-consistent $(n, p \rightarrow \infty)$ estimator $\tilde{F}$ of F^* , under what conditions and at what rate will $\tilde{\Sigma} = f^{-1}(\tilde{F})$ converge to $\Sigma^* = f^{-1}(F^*)$ in the spectral norm?

2.2. Elementwise perturbation theory for matrix transformations

Theorem 1 answers the first of the two questions from the previous paragraph in general terms. The result hinges on the following condition.

Condition 1. As $p \rightarrow \infty$, the sequence of minimum eigenvalues of the corresponding matrices Σ^* is bounded away from zero.

In the case of a diagonal covariance matrix, we require Σ^* to have minimum diagonal entry bounded away from zero as $p \rightarrow \infty$, and Condition 1 can be viewed as a generalization of that. A lower bound on λ_p^* is $\{\sum_{k=1}^p (\lambda_k^*)^{-2}\}^{-1/2} = \text{tr}(\Omega^{*2})^{-1/2}$ so that a sufficient requirement for Condition 1 is that the sum of the diagonal entries of the squared concentration matrix be bounded, although it could diverge with p .

THEOREM 1. Suppose that Condition 1 holds. Let $\hat{\Sigma}^+$ be a symmetric matrix with ordered eigenvalues $\hat{\lambda}_1, \dots, \hat{\lambda}_p$ where, as $p \rightarrow \infty$, the sequence of smallest eigenvalues $\hat{\lambda}_p$ is bounded away from zero. Let $\hat{F}^P = f(\hat{\Sigma}^+)$. Then $\hat{F}^P$ is elementwise consistent $(n, p \rightarrow \infty)$ for F^* if $\hat{\Sigma}^+$ is elementwise consistent for Σ^* as $n, p \rightarrow \infty$ and

$$\text{Res} \left\{ \frac{f(z)}{(z - \hat{\lambda}_k)(z - \lambda_v^*)}, \hat{\lambda}_k \right\} + \text{Res} \left\{ \frac{f(z)}{(z - \hat{\lambda}_k)(z - \lambda_v^*)}, \lambda_v^* \right\}, \quad z \in \mathbb{C}, \quad (2)$$

is bounded in absolute value for all $1 \leq k, v \leq p$ as $n, p \rightarrow \infty$, where, for a function g of a complex variable, $\text{Res}(g, a)$ denotes the residue of g at a . If this condition is satisfied, then $\|\hat{F}^P - F^*\|_{\max} \asymp \|\hat{\Sigma}^+ - \Sigma^*\|_{\max}$.

The conclusion $\|\hat{F}^P - F^*\|_{\max} \asymp \|\hat{\Sigma}^+ - \Sigma^*\|_{\max}$ is true even when $\hat{\Sigma}^+$ is not elementwise-consistent as $n, p \rightarrow \infty$, but this implies that $\hat{F}^P$ is not elementwise consistent either. Rates of convergence of elementwise-consistent estimators of Σ^* are discussed in § 2.4.

Some important special cases of Theorem 1 are discussed in less abstract terminology after the proof. We henceforth write the sum of the residues from (2) as $R_{k,v}$.

Proof. A function f of a $p \times p$ matrix A satisfies (Kato, 1976, p. 44)

$$f(A) = \frac{1}{2\pi i} \oint_{\gamma_A} f(z)(zI - A)^{-1} dz,$$

where I is the identity matrix and γ_A is a simple closed curve lying in the region of analyticity of f and enclosing all the eigenvalues of A in its interior. Then, in an obvious notation,

$$\hat{F}^P - F^* = \frac{1}{2\pi i} \left\{ \oint_{\hat{\gamma}} f(z)(zI - \hat{\Sigma}^+)^{-1} dz - \oint_{\gamma^*} f(z)(zI - \Sigma^*)^{-1} dz \right\},$$

and the deformation theorem (Whittaker & Watson, 1920, § 5.2, Corollary 2), ensures that for $\hat{\gamma}$ and γ^* lying wholly in γ , which does not cross the imaginary axis,

$$\begin{aligned} \hat{F}^P - F^* &= \frac{1}{2\pi i} \oint_{\gamma} f(z) \{ (zI - \hat{\Sigma}^+)^{-1} - (zI - \Sigma^*)^{-1} \} dz \\ &= \frac{1}{2\pi i} \oint_{\gamma} f(z)(zI - \hat{\Sigma}^+)^{-1} (\hat{\Sigma}^+ - \Sigma^*)(zI - \Sigma^*)^{-1} dz, \end{aligned}$$

where we have used the fact that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for invertible matrices A and B . Writing a positive-definite matrix A as QDQ^T where D is A 's diagonal matrix of eigenvalues and Q is the matrix of corresponding orthonormal eigenvectors, $(zI - A)^{-1} = Q(zI - D)^{-1}Q^T$, so

$$\|\hat{F}^P - F^*\|_{\max} = \max_{q,j} \left| \frac{1}{2\pi i} \sum_{k,v,\ell,s} \left\{ \oint_{\gamma} \frac{f(z)}{(z - \hat{\lambda}_k)(z - \lambda_v^*)} dz \right\} \hat{\xi}_{qk} \hat{\xi}_{\ell k} \xi_{sv}^* \xi_{jv}^* (\hat{\Sigma}_{\ell s}^+ - \Sigma_{\ell s}^*) \right| \quad (3)$$

and, by the theory of residues (Whittaker & Watson, 1920, p. 112),

$$\frac{1}{2\pi i} \oint_{\gamma} \frac{f(z)}{(z - \hat{\lambda}_k)(z - \lambda_v^*)} dz = R_{k,v}.$$

The orthonormality of the eigenvectors ensures that their entries are of order $1/p$ on average. Thus the quadruple sum in (3) causes no accumulation of error as p diverges, and the convergence rate of $f(\hat{\Sigma}^+)$ to $f(\Sigma^*)$ is that of $\hat{\Sigma}^+$ to Σ^* , provided that the $R_{k,v}$ are bounded in absolute value for all $1 \leq k, v \leq p$ as $n, p \rightarrow \infty$. $\square$

COROLLARY 1. Let $\hat{\Sigma}^+$ be as in Theorem 1, and let $\hat{L}^P = \log(\hat{\Sigma}^+)$. Under Condition 1, $\|\hat{L}^P - L^*\|_{\max} \asymp \|\hat{\Sigma}^+ - \Sigma^*\|_{\max}$, where $L^* = \log(\Sigma^*)$.

Proof. By direct calculation

$$\frac{1}{2\pi i} \oint_{\gamma} \frac{\log z}{(z - \hat{\lambda}_k)(z - \lambda_v^*)} dz = R_{k,v} = \frac{\log \hat{\lambda}_k - \log \lambda_v^*}{\hat{\lambda}_k - \lambda_v^*}, \quad (4)$$

which is unbounded only in the limit of λ_v^* or $\hat{\lambda}_k$ tending to zero, which is precluded by the conditions of Corollary 1. The limit as $\hat{\lambda}_k \rightarrow \lambda_v^*$ is λ_v^{*-1} , which is bounded as $p \rightarrow \infty$ by Condition 1. $\square$

The remarks about convergence rates immediately following Theorem 1 apply. The special case of the matrix logarithm makes it clear that the conclusion of, and proof strategy for, Theorem 1 are far from evident a priori in the $p \geq n$ setting. While the simplest proof of Corollary 1 for $p < n$ would show consistency of the eigenvectors and log eigenvalues of $\hat{\Sigma}^+$ for those of Σ^* as $n \rightarrow \infty$, this is obviously unsuitable when $p > n$, by the discussion of § 1. Since we are only interested in a neighbourhood of the spectra of Σ^* and $\hat{\Sigma}^+$, we can select for $\log z$ in the proof of Corollary 1 any branch of the complex logarithm and have no difficulties due to its multivaluedness.

Remark 1. By a different strategy, Birman & Solomyak (2003, Theorem 8.1) proved that for general self-adjoint operators A and B , and for functions f that are uniformly Lipschitz on $\mathbb{R}$,

$$f(B) - f(A) = Z_f^{A,B}(B - A), \quad (5)$$

where

$$Z_f^{A,B} = \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{f(\mu) - f(\zeta)}{\mu - \zeta} dE^A(\zeta) dF^B(\mu)$$

with E^A and F^B being the spectral measures of A and B ; this rules out the case of the logarithm and the inverse. With discrete spectral measures, (5) is similar to (3) with its contour integral replaced by the right-hand side of (4).

Another important example of F^* is the matrix inverse of Σ^* . For the interpretation of elements of a concentration matrix, see Cox & Wermuth (1996, pp. 68–9).

COROLLARY 2. Let $\hat{\Sigma}^+$ be as in Theorem 1, and let $\hat{\Omega}^P = (\hat{\Sigma}^+)^{-1}$. Under Condition 1, $\|\hat{\Omega}^P - \Omega^*\|_{\max} \asymp \|\hat{\Sigma}^+ - \Sigma^*\|_{\max}$.

Proof. The proof is, mutatis mutandis, that of Corollary 1. In particular,

$$R_{k,v} = \{\lambda_v^*(\lambda_v^* - \hat{\lambda}_k)\}^{-1} + \{\hat{\lambda}_k(\hat{\lambda}_k - \lambda_v^*)\}^{-1} = -(\hat{\lambda}_k \lambda_v^*)^{-1},$$

which is bounded as $n, p \rightarrow \infty$ by Condition 1 and the specification of $\hat{\Sigma}^+$. $\square$

The implication of this result is that a constrained optimization-based approach (e.g., Cai et al., 2016) is unnecessary for estimating a sparse matrix inverse. Consistency as $n, p \rightarrow \infty$ is achieved by thresholding $\hat{\Omega}^P$, as shown in Corollary 3 below. This has appreciable computational benefits.

2.3. Consistent estimation of covariance and concentration matrices in the spectral norm

As discussed in § 1, success in many applications depends on covariance or concentration matrix estimators that are consistent ($n, p \rightarrow \infty$) in the spectral norm. Such consistency is achievable under sparsity. Accordingly, we assume that a transformation f has been found for which the following holds.

Condition 2. The matrix function $F^* = f(\Sigma^*) = (F_{uv}^*)$ belongs to the class of matrices

$$\mathcal{F}\{s(p), q\} = \left\{ F^* : F_{uv}^* = F_{vu}^*, \max_u |F_{uu}^*| \leq C, \max_u \sum_{v=1}^p |F_{uv}^*|^q \leq s(p) \right\},$$

where C is an arbitrary real constant, $0 \leq q \leq 1$, and $s(p)/p \rightarrow 0$.

Define a thresholding estimator of F^* as $\tilde{F} = (\tilde{F}_{uv})$ where

$$\tilde{F}_{uv} = \mathcal{T}_\nu(\hat{F}_{uv}^P) = \hat{F}_{uv}^P \mathbb{I}(|\hat{F}_{uv}^P| > \nu) \quad (6)$$

for a pilot estimator $\hat{F}^P$ and an appropriate choice of threshold ν to be discussed below. The next result follows immediately from Theorem 1 and the literature on thresholded estimators (e.g., Bickel & Levina, 2008, proof of Theorem 1).

COROLLARY 3. Let $\hat{\Sigma}^+$ be as in Theorem 1 and such that $\|\hat{\Sigma}^+ - \Sigma^*\|_{\max} = O_{\text{pr}}(r_{n,p})$, where $r_{n,p}$ tends to zero as $n, p \rightarrow \infty$. Assume Conditions 1 and 2 and take $\nu \asymp r_{n,p}$ in (6). Then $\|\tilde{F} - F^*\|_{\text{op}} = O_{\text{pr}}\{s(p)r_{n,p}^{1-q}\}$ as $n, p \rightarrow \infty$.

Remark 2. The requirement that $r_{n,p} \rightarrow 0$ as $n, p \rightarrow \infty$ usually implies the scaling condition $(\log p)/n \rightarrow 0$. Further details follow Corollary 4.

While Corollary 3 describes the rate at which the tuning parameter may decay as p and n increase, in small samples the constant c in this tuning parameter is relevant and asymptotic results provide no guidance on how to set it. All such sparsity-inducing procedures suffer this limitation. The simplest strategy is to choose a reasonable arbitrary constant such as $c = 1$ so that, for the pilot estimators considered in § 2.4, one would set $r_{n,p} = \{(\log p)/n\}^{1/2}$ under appropriate tail conditions, although one may set c using a measure of predictive performance out of sample. The construction of $\hat{\Sigma}^+$ given in Proposition 2 makes calculation of $\tilde{F}$ almost instantaneous for $p \leq 1000$, facilitating crossvalidation. It is reasonable in practice, although statistically suboptimal, to decide on a level α of sparsity that one is willing to impose and set the threshold so that only the proportion α of entries that are largest in absolute value are kept. This was done in most of the experiments of § 4. One might instead inspect a plot of the number of matrix entries above ν against ν , and choose a value of ν at which a sharp decrease in negative gradient, an elbow, is observed, if any.

Given a spectral norm-consistent estimator $\tilde{F}$ of F^* as $n, p \rightarrow \infty$, the remaining challenge is to establish conditions under which $f^{-1}(\tilde{F})$ converges to $f^{-1}(F^*)$ in the spectral norm, with associated convergence rates. Under smoothness conditions on $g = f^{-1}$, and for symmetric matrices A and B , general bounds of the form

$$\|g(A) - g(B)\|_{\text{op}} \leq c_g \|g'\|_{\infty} \|A - B\|_{\text{op}} \quad (7)$$

are available. In (7), c_g , is a constant, g' is the derivative of g , and $\|g'\|_\infty = \sup_x |g'(x)|$, where the matrix-valued function is defined as the same scalar-valued function applied to the eigenvalues. See Bhatia (1997, p. 320) for references and Bhatia (1997, § X) for other bounds under special conditions on g, A and B . Equation (7) is, however, useless for unbounded functions, which are best studied on a case-by-case basis as in Theorem 2.

THEOREM 2. Assume Conditions 1 and 2 and let $\tilde{\Sigma} = \exp(\tilde{L})$ and $\tilde{\Omega} = \exp(-\tilde{L})$, where $\tilde{L}$ is the thresholded pilot estimator of $L^* = \log(\Sigma^*)$, i.e., $\tilde{L} = \{\mathcal{T}_v(\hat{L}_{uv}^P)\}$, where $\hat{L}^P = \log(\hat{\Sigma}^+)$ and v and $\hat{\Sigma}^+$ are as in Corollary 3. Then under the conditions of Corollary 3,

$$\begin{aligned}\|\tilde{\Sigma} - \Sigma^*\|_{\text{op}} &\leq \exp\{|\lambda_1(L^*)|\} \|\tilde{L} - L^*\|_{\text{op}} \exp(\|\tilde{L} - L^*\|_{\text{op}}), \\ \|\tilde{\Omega} - \Omega^*\|_{\text{op}} &\leq \exp\{|\lambda_p(L^*)|\} \|\tilde{L} - L^*\|_{\text{op}} \exp(\|\tilde{L} - L^*\|_{\text{op}}).\end{aligned}$$

Thus if the sequences of maximum and minimum eigenvalues of L^* are bounded as $p \rightarrow \infty$, $\tilde{\Sigma}$ and $\tilde{\Omega}$ are consistent ($n, p \rightarrow \infty$) for Σ^* and Ω^* in the spectral norm as $n, p \rightarrow \infty$. They are also positive definite by construction.

2.4. Pilot estimators for Σ^*

We have not discussed construction of elementwise-consistent pilot estimators $\hat{\Sigma}^+$ as $n, p \rightarrow \infty$ that have minimum eigenvalues bounded away from zero, which is crucial for application of Theorem 1. A natural choice is the projection, in elementwise loss, of an unconstrained estimator $\hat{\Sigma}^P$ onto the cone $\{\Sigma : \Sigma = \Sigma^T, \Sigma - \delta I_p \geq 0\}$, where $\delta > 0$ is discussed below in §2.4. The projection,

$$\hat{\Sigma}^+ = \arg \min_{\Sigma \geq \delta I_p} \|\hat{\Sigma}^P - \Sigma\|_{\max}, \quad (8)$$

can be rewritten as the semidefinite programming problem

$$\min_{\vartheta > 0, \Sigma - \delta I_p \geq 0} \vartheta : |(\hat{\Sigma}^P - \Sigma)_{uv}| \leq \vartheta, \quad u, v \in [p],$$

and solved using the `cvx` solver in Matlab. The projection preserves the elementwise convergence rate of $\hat{\Sigma}^P$, as shown in Proposition 1.

PROPOSITION 1. If $\delta \leq \lambda_p^*$ in (8), then $\|\hat{\Sigma}^+ - \Sigma^*\|_{\max} \leq 2\|\hat{\Sigma}^P - \Sigma^*\|_{\max}$.

In the case of the matrix logarithm, the largest contribution to the finite-sample elementwise error in estimating L^* by $\hat{L}^P$ is seen from (3) and (4) to be of order

$$\|\hat{\Sigma}^+ - \Sigma^*\|_{\max} \frac{\log \delta - \log \lambda_p^*}{\delta - \lambda_p^*}.$$

Thus, with the constraint that $\delta \leq \lambda_p^*$, which ensures that $\|\hat{\Sigma}^+ - \Sigma^*\|_{\max} \asymp \|\hat{\Sigma}^P - \Sigma^*\|_{\max}$, the finite-sample error is minimized in the limit as $\delta \rightarrow \lambda_p^*$. In the absence of information about λ_p^* , we are forced to choose a small arbitrary constant. Another construction of an elementwise-consistent $\hat{\Sigma}^+$ is in Proposition 2.

PROPOSITION 2. Suppose $\|\hat{\Sigma}^P - \Sigma^*\|_{\max} = O_{\text{pr}}(q_{n,p})$ and define $\hat{\Sigma}^+$ as the convex combination $(1 - \delta_{n,p})\hat{\Sigma}^P + \delta_{n,p}\hat{D}$ where $\hat{D} = \text{diag}(\hat{\Sigma}^P)$ and $\delta_{n,p} \in (0, 1)$ tends to zero as $n, p \rightarrow \infty$. Then $\|\hat{\Sigma}^+ - \Sigma^*\|_{\max} = O_{\text{pr}}\{\max(q_{n,p}, \delta_{n,p})\}$.

This $\hat{\Sigma}^+$ is a biased estimator of Σ^* in finite samples but asymptotically unbiased by the requirement on $\delta_{n,p}$ as $n, p \rightarrow \infty$. The following corollary of Weyl's inequality elucidates the effect of $\delta_{n,p}$.

COROLLARY 4 (Bhatia 1997, Corollary III.2.3). Let A and B be $p \times p$ symmetric matrices. For every $j = 1, \dots, p$,

$$\lambda_j(A) + \lambda_p(B) \leq \lambda_j(A + B) \leq \lambda_j(A) + \lambda_1(B).$$

The implication in the present context is that, with $\hat{\Sigma}^+$ as in Proposition 2, $\delta_{n,p}\lambda_p(\hat{D}) \leq \lambda_p(\hat{\Sigma}^+)$ and $\lambda_1(\hat{\Sigma}^+) \leq (1 - \delta_{n,p})\lambda_1(\hat{\Sigma}^P) + \delta_{n,p}\lambda_1(\hat{D})$. Therefore, $\hat{\Sigma}^P \mapsto (1 - \delta_{n,p})\hat{\Sigma}^P + \delta_{n,p}\hat{D}$ pulls the lower eigenvalue away from zero but may also increase the upper eigenvalues.

If $\hat{\Sigma}^P$ is the sample covariance matrix, then $\|\hat{\Sigma}^P - \Sigma^*\|_{\max} = O_{\text{pr}}[\{(\log p)/n\}^{1/2}]$ assuming sub-Gaussianity of $X_1, \dots, X_n$. Other choices of pilot estimator achieve the same rate under less stringent tail conditions; see Avella-Medina et al. (2018).

The following section describes the implications of the previous results for high-dimensional linear discriminant analysis and linear regression, in which a covariance or concentration matrix is a nuisance parameter. These examples are used in § 4 to assess empirical performance.

3. IMPLICATIONS

3.1. Discriminant analysis

One frequently encountered high-dimensional data type involves binary labelled data, $y_i \in \{-1, 1\}$, alongside high-dimensional covariate information x_i .

Discriminant analysis aims to find a function $D(x)$ of the features that achieves maximal separability between the two classes $\mathcal{C}_+ = \{i : y_i = 1\}$ and $\mathcal{C}_- = \{i : y_i = -1\}$, and one way of quantifying separability is the classification risk,

$$\text{CR}(D) = \kappa P_X^+[\mathbb{I}\{D(X) < 0\}] + (1 - \kappa)P_X^-[\mathbb{I}\{D(X) > 0\}],$$

where $\kappa = \text{pr}(Y = 1)$ and $P_X^\pm$ is the conditional distribution of X given its class label, $+$ or $-$. Fisher (1936) suggested to quantify separability using

$$\text{RF}(D) = \frac{\text{var}[E\{D(X) \mid Y\}]}{\text{var}[\{D(X) \mid Y\} - E\{D(X) \mid Y\}]},$$

the ratio of between-class variability to within-class variability. Thus maximization of $\text{RF}(D)$ with respect to D corresponds to finding the discriminant function such that the proportion of variability explained by the class labels is as large as possible. Although minimization of $\text{CR}(D)$ requires knowledge of the conditional distributions $P_X^\pm$, maximization of $\text{RF}(D)$ only requires knowledge, or estimation, of moments.

For linear discriminant rules, $D(X) = w^T X$ and $\text{RF}(D) = \text{RF}(w) = \kappa(1 - \kappa)R(w)$, where

$$R(w) = \frac{\{E(w^T X \mid Y = -1) - E(w^T X \mid Y = 1)\}^2}{\kappa \text{var}(w^T X \mid Y = 1) + (1 - \kappa) \text{var}(w^T X \mid Y = -1)} = \frac{\{w^T(\mu_-^* - \mu_+^*)\}^2}{w^T\{\kappa \Sigma_+^* + (1 - \kappa) \Sigma_-^*\}w}, \quad (9)$$

with μ_-^* , μ_+^* , Σ_-^* and Σ_+^* being the means and covariance matrices of $X \mid Y = \pm 1$.

The Supplementary Material gives an adaptation of the Fisher (1936) procedure suitable when $p \gg n$ and establishes its theoretical properties when a spectral norm-consistent ($n, p \rightarrow \infty$) estimator of Σ^* is available. The procedure relies on convex optimization and a simpler strategy exploits the connection to linear regression, indicated by Fisher (1936) and discussed in § 4.3.

3.2. Linear regression analysis

Regression coefficients express the dependence of a response variable Y on explanatory variables $X_1, \dots, X_p$ and, in linear relationships, have the simple interpretation of giving the change in $E(Y)$ for a hypothetical unit increase in X_s , say, with the values of the other variables held fixed. Linear regression coefficients are of the form

$$\beta^* = (\Sigma_{YX}^* \Omega_{XX}^*)^T$$

where Σ_{YX}^* is the row vector of elements $\text{cov}(Y, X_s)$ and $\Omega_{XX}^* = (\Sigma_{XX}^*)^{-1}$, with Σ_{XX}^* being the matrix of covariances between $X_1, \dots, X_p$.

THEOREM 3. *Let $\hat{\Sigma}$ be an estimator of Σ^* , the covariance matrix of $(Y, X_1, \dots, X_p)$, and let $\hat{\Omega}_{XX}$ be an estimator of Ω_{XX}^* . Define the regression coefficient estimator as $\hat{\beta} = (\hat{\Sigma}_{YX} \hat{\Omega}_{XX})^T$ where $\hat{\Sigma}_{YX}$ is the first column of $\hat{\Sigma}$. Then*

$$\|\hat{\beta} - \beta^*\|_2 \leq \|\hat{\Sigma} - \Sigma^*\|_{\text{op}} \|\hat{\Omega}_{XX}\|_{\text{op}} + \|\Sigma_{YX}^*\|_2 \|\hat{\Omega}_{XX} - \Omega_{XX}^*\|_{\text{op}}.$$

Under regularity conditions, a strong form of consistency of $\hat{\beta}$ as $n, p \rightarrow \infty$ is guaranteed by estimators of Σ^* and Ω_{XX}^* that are consistent in the spectral norm, as constructed in § 2 assuming the existence of an approximate sparsity-inducing transformation.

4. EXAMPLES

4.1. Idealized example

The performance of the proposed procedure under a logarithmic sparsity assumption is investigated as follows. Generate n replicates of $p = 200$ variables from a normal distribution of mean zero and covariance matrix Σ^* , where n takes values $\{20, 60, \dots, 180\}$ and $\Sigma^* = \exp(L^*)$, with $L^* = \sum_{m=1}^{|\mathcal{B}|} \alpha_m^* B_m$ for matrices $B_m \in \mathcal{B}$, the natural symmetrized basis for the space of symmetric matrices. Specifically, $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$, where $\mathcal{B}_1$ consists of p diagonal indicator components of the form

$$\mathcal{B}_1 = \{B : B = e_j e_j^T, j \in [p]\}$$

and $\mathcal{B}_2$ consists of $p(p-1)/2$ symmetric nondiagonal indicator components of the form

$$\mathcal{B}_2 = \{B : B = e_j e_k^T + e_k e_j^T, j, k \in [p], j \neq k\}.$$

The support of α^* is taken as a simple random sample of size 60 from the index set $[p(p+1)/2]$. Half of the nonzero entries of α^* are generated from a uniform distribution on $[0.5, 1]$ and the other half from a uniform distribution on $[-1, -0.5]$. This realization of α^* is then fixed across 100 Monte Carlo replications. The resulting L^* belongs to the class $\mathcal{F}\{s(p), q\}$ with $q = 0$.

Table 1 shows the performance in the spectral and Frobenius norms of four estimators of Σ^* : the sample covariance and its thresholded version, and the pilot estimator $\hat{\Sigma}^+$ and the matrix

Table 1. *Relative estimation errors ($\times 100$) in the spectral and Frobenius norms for covariance and concentration matrix estimation*

n	Error	Estimators E of Σ^*				Error	Estimators E of Ω^*		
		S_n	$\mathcal{T}(S_n)$	$\hat{\Sigma}^+$	$\tilde{\Sigma}$		$\hat{\Omega}^C$	$(\hat{\Sigma}^+)^{-1}$	$\tilde{\Omega}$
20	$\ E - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	255	208	107	92.9	$\ E - \Omega^*\ _{\text{op}}/\ \Omega^*\ _{\text{op}}$	99.3	86.20	92.9
20	$\ E - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	200	171	106	83.5	$\ E - \Omega^*\ _{\text{F}}/\ \Omega^*\ _{\text{F}}$	95.2	77.28	83.8
60	$\ E - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	121	100	74.1	71.8	$\ E - \Omega^*\ _{\text{op}}/\ \Omega^*\ _{\text{op}}$	98.1	76.56	72.3
60	$\ E - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	115	98.3	84.4	63.2	$\ E - \Omega^*\ _{\text{F}}/\ \Omega^*\ _{\text{F}}$	92.3	79.21	64.4
100	$\ E - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	86.9	72.0	59.1	58.5	$\ E - \Omega^*\ _{\text{op}}/\ \Omega^*\ _{\text{op}}$	95.3	70.71	61.2
100	$\ E - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	89.0	76.3	71.2	48.4	$\ E - \Omega^*\ _{\text{F}}/\ \Omega^*\ _{\text{F}}$	87.8	79.89	51.2
140	$\ E - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	70.8	58.8	50.6	51.3	$\ E - \Omega^*\ _{\text{op}}/\ \Omega^*\ _{\text{op}}$	87.9	66.68	55.9
140	$\ E - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	75.1	64.4	62.6	40.8	$\ E - \Omega^*\ _{\text{F}}/\ \Omega^*\ _{\text{F}}$	82.4	76.34	44.7
180	$\ E - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	61.6	51.3	45.8	46.6	$\ E - \Omega^*\ _{\text{op}}/\ \Omega^*\ _{\text{op}}$	85.0	63.72	51.6
180	$\ E - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	66.3	56.9	56.8	36.3	$\ E - \Omega^*\ _{\text{F}}/\ \Omega^*\ _{\text{F}}$	78.4	70.84	40.3
Largest standard error		29	31	13	3.7		1.4	0.57	2.8

Average estimation errors over 100 Monte Carlo replications for covariance matrix estimation, columns 3–6, and concentration matrix estimation, columns 8–10; rounding errors are less than standard errors divided by 10 and the largest standard error for each column is reported at the foot.

exponential of the thresholded matrix logarithm of $\hat{\Sigma}^+$, namely $\tilde{\Sigma} = \exp(\tilde{L})$ where $\tilde{L} = \mathcal{T}_v(\hat{L}^P)$. The pilot estimator $\hat{\Sigma}^+$ used here is the shrinkage-based pilot estimator with shrinkage parameter $\delta_{n,p} = \{(\log p)/n\}^{1/2}$. For thresholding in the matrix logarithmic domain, the tuning parameter is $\nu = \{(\log p)/n\}^{1/2}$. That is, a tuning constant $c = 1$ is chosen in both stages. For thresholding in the original domain the tuning parameter is taken to be $\nu = c\{(\log p)/n\}^{1/2}$, with c the value between 0.1 and 3 that minimizes the error. All errors are normalized by $\|\Sigma^*\|$, where the norm is that in which the errors are measured.

For the same Σ^* in every experiment, we also consider estimation of $\Omega^* = \Sigma^{*-1}$. The inverse sample covariance matrix does not exist for $p \geq n$ and that of the thresholding estimator may not exist either, so we compare $\tilde{\Omega} = \exp(-\tilde{L})$ to the inverse of $\hat{\Sigma}^+$ and to $\hat{\Omega}^C$, the adaptively constrained ℓ_1 minimization estimator of Cai et al. (2016). Calculation of $\hat{\Omega}^C$ involves $2p$ p -dimensional optimization problems, each with more than $4p$ constraints, which is computationally expensive and limits the extent of the testing. The tuning parameter for $\hat{\Omega}^C$ is taken as $c\{(\log p)/n\}^{1/2}$. In the current experiments $c = 1$ and $c = 2$ are tested, the latter being the value suggested by Cai et al. (2016). The results for $c = 1$ are reported, as these are better than with $c = 2$. The relative estimation errors in the spectral and Frobenius norms are shown in Table 1.

Computation of the matrix logarithm and exponential in high dimensions should not be attempted using the spectral representation. A more accurate method by Padé approximation is implemented in the `logm` and `expm` functions of Matlab.

4.2. Misspecified sparsity structure

We illustrate here that there can be advantages in using a sparsity scale different from the original when no structure is present.

We assess through simulations the approximation and estimation errors from assuming that Σ^* possesses sparsity structure when, in fact, it possesses none, showing empirically that wrongly assuming matrix logarithmic sparsity is roughly equivalent to wrongly assuming sparsity of Σ^* in terms of approximation error but, for the situation considered here, is better in terms of estimation error when $p \gg n$. It is also better than making no sparsity assumptions. By approximation error, we mean the error incurred by replacing the true object with one in which sparsity is imposed on

Table 2. *Relative approximation and estimation errors ($\times 100$) in the spectral and Frobenius norms when no sparsity structure is present*

(τ, κ)	Error	Matrix approximation or estimator (M)				
		$\mathcal{T}(\Sigma^*)$	$\exp\{\mathcal{T}(L^*)\}$	S_n	$\mathcal{T}(S_n)$	$\exp\{\mathcal{T}(\hat{L}^P)\}$
(2,2)	$\ M - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	44.8 (0.76)	64.5 (2.5)	366.4 (48)	174.5 (23)	85.95 (1.3)
(2,2)	$\ M - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	43.3 (0.83)	47.7 (0.96)	258.1 (3.0)	176.7 (1.8)	75.67 (0.33)
(4,2)	$\ M - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	40.4 (0.98)	54.1 (2.3)	494.0 (49)	239.5 (24)	83.80 (0.99)
(4,2)	$\ M - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	33.5 (0.63)	35.3 (0.70)	283.2 (2.0)	190.8 (1.3)	72.62 (0.19)
(2,4)	$\ M - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	44.7 (0.81)	64.6 (2.8)	363.1 (53)	173.6 (26)	85.64 (1.0)
(2,4)	$\ M - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	43.3 (0.80)	47.6 (0.94)	258.3 (3.0)	176.8 (1.8)	75.64 (0.31)
(4,4)	$\ M - \Sigma^*\ _{\text{op}}/\ \Sigma^*\ _{\text{op}}$	40.4 (1.0)	54.0 (2.3)	495.8 (51)	240.9 (25)	83.65 (1.1)
(4,4)	$\ M - \Sigma^*\ _{\text{F}}/\ \Sigma^*\ _{\text{F}}$	33.6 (0.61)	35.4 (0.66)	282.7 (1.9)	190.5 (1.2)	72.65 (0.20)

Average (standard deviation in parentheses) over 100 Monte Carlo replications of: approximation errors, columns 3 and 4, from wrongly assuming sparsity in the original domain and in the logarithmic domain; estimation error of the sample covariance matrix, column 5, which correctly assumes no structure and estimation errors from wrongly assuming sparsity in the original domain and in the logarithmic domain, columns 6 and 7.

the appropriate scale, once transformed back to the original scale. By estimation error, we mean the error incurred by replacing Σ^* with an estimator.

The experiment is as follows. A matrix Ξ of orthonormal eigenvectors is constructed by generating a $p \times p$ matrix Z whose entries are independent and standard normal and computing the QR -factorization $Z = QR$. Then Ξ is set equal to Q . This matrix is fixed, and in each of 100 Monte Carlo replications $p = 1000$ eigenvalues are generated at random from a gamma distribution of shape parameter τ and scale parameter κ , varied as in Table 2. A random unstructured covariance matrix is then constructed as $\Sigma^* = \Xi \Lambda \Xi^T$ where Λ is a diagonal matrix containing the gamma-distributed eigenvalues. Both Σ^* and $L^* = \log(\Sigma^*)$ are thresholded so that only the entries above the 0.95 quantile in absolute value are kept. The resulting approximations, sparse in the original and logarithmic domains respectively, are denoted by $\mathcal{T}(\Sigma^*)$ and $\exp\{\mathcal{T}(L^*)\}$ in Table 2.

A random sample of size $n = p/10 = 100$ is drawn from a normal distribution of mean zero and covariance matrix Σ^* , and the sample covariance S_n is constructed along with $\hat{\Sigma}^+ = (1 - \delta_{n,p})S_n + \delta_{n,p}\text{diag}(S_n)$ where $\delta_{n,p} = \{\log(p)/n\}^{1/2}$. The estimation errors of S_n , its thresholded version and the matrix exponential of the thresholded value of $\hat{L}^P = \log(\hat{\Sigma}^+)$ are reported in Table 2. Thresholding of both matrices is such that only entries above the 0.95 quantile in absolute value are kept.

It is unclear why an erroneous assumption of log sparsity can lead to worse approximation error but better estimation error than an erroneous assumption of sparsity in the original domain. However, the logarithmic transformation lessens the contributions of the large eigenvalues to estimation of the sparsity structure. Large eigenvalues are shrunk relatively more than small ones. This seems sensible since the large eigenvalues of $\hat{\Sigma}^P$ and hence $\hat{\Sigma}^+$ are severely overdispersed. The lower eigenvalues have already been regularized by conversion of $\hat{\Sigma}^P$ to $\hat{\Sigma}^+$.

A further simulation study, showing the effect of assuming matrix logarithmic sparsity of Σ^* when it instead satisfies Condition 2 in the original domain, is in the Supplementary Material.

4.3. The Fisher procedure with misspecified covariance structure

While the goal of discriminant analysis is understanding of the data-generating process rather than prediction, performance in terms of the latter provides a way of assessing the proposed

Table 3. *Out-of-sample misclassification rate ($\times 100$) of an adaptation of linear discriminant analysis and a high-dimensional classifier*

	(N_+, N_-)				
	(20, 30)	(40, 50)	(60, 60)	(100, 100)	(200, 200)
Sparse L^* regression	16 (8.3)	13 (7.9)	11 (7.2)	8 (5.9)	7 (5.3)
High-dimensional classifier	32 (12)	24 (11)	18 (10)	11 (7.7)	6 (5.3)

Average (standard deviation in parentheses) over 100 Monte Carlo replications; N_+ and N_- are the numbers of observations in the training sample for each class, and the number of features is $p = 500$.

procedure. We calculate the mean and standard deviation of the missclassification rate from out-of-sample classification. In each of 100 Monte Carlo replications, N_+ realizations of $p = 500$ normally distributed features of mean μ_+ and covariance Σ_+ are obtained together with N_- realizations of p normally distributed features of mean μ_- and covariance Σ_- . Call the corresponding $N_+ \times p$ and $N_- \times p$ matrices x_+ and x_- . The entries of the two mean vectors are standard uniformly distributed, and the two covariance matrices are generated as in the first example in § 4.2 with $(\tau, \kappa) = (2, 2)$. The two mean vectors and covariance matrices are drawn anew on each replication.

As pointed out by Fisher (1936), discriminant analysis as described in § 3.1 for p much smaller than the sample size is equivalent to using as the linear discriminant the coefficient vector from a linear regression in which the class labels are viewed as response variables and assigned the values $Y_+ = N_-/(N_+ + N_-)$ for class $+$ and $Y_- = -N_+/(N_+ + N_-)$ for class $-$, while the multivariate normally distributed data for each class are viewed as explanatory variables.

When p is much larger than $(N_+ + N_-)$, which is the situation here, standard linear regression is not feasible, so we adapt it as outlined in § 3.2 using $\tilde{\Omega}_{XX} = \exp(-\tilde{L})$ where $\tilde{L}$ is the 95% thresholded version of $\hat{L}^P$, the latter constructed as in § 4.2 from the sample covariance matrix of x , the matrix whose first N_+ rows are x_+ and whose remaining rows are x_- . In each replication a separate test set is generated, consisting of 10 observations from the $+$ class and 10 observations from the $-$ class. The new individuals are classified into $+$ and $-$ based on the signs of their linear predictors and the missclassification rates recorded. Table 3 gives the results for different values of N_+ and N_- . Since the procedures of the present paper were not designed for classification, it is not expected that the suggested modification of the Fisher (1936) procedure outperforms other high-dimensional classifiers, but it seems competitive. Table 3 reports the missclassification rates of another high-dimensional classifier (Fan et al., 2012), for which the tuning parameters are chosen by crossvalidation using the author's codes, available at <https://uk.mathworks.com/matlabcentral/fileexchange/40047>.

4.4. A real example

We apply the modification of the Fisher (1936) procedure as described in § 4.3 to the leukaemia data used by Efron & Hastie (2016, § 19.1). The data consist of $p = 3571$ genetic features for 72 patients, of whom 47 have acute lymphoblastic leukaemia and 25 have acute myeloid leukaemia, a worse prognosis. We remove each patient from the dataset, use the data on the remaining patients to estimate the linear discriminant as described in § 4.3 with the $\hat{\Sigma}^+$ of § 4.2, and predict the excluded patient's class label using his or her covariate information and the estimated coefficient vector. The missclassification rate is 8/72, much better than random guessing and possibly improvable with a data-dependent tuning parameter. The high-dimensional classification procedure of Fan et al. (2012) with tuning parameters chosen by crossvalidation produced 6/72 misclassifications.

ACKNOWLEDGEMENT

Sincere thanks are due to Nancy Reid, whose comments changed the bearing of the work, and to the referees and editors for helpful criticism. I benefitted from relevant discussions with Marco Avella-Medina, David Cox, and Jianqing Fan. The work was partially supported by the U.K. Engineering & Physical Sciences Research Council.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online shows implications of the results for several applications and includes proofs omitted from the paper and a further simulation study.

REFERENCES

- AVELLA-MEDINA, M., BATTEY, H. S., FAN, J. & LI, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika* **105**, 271–84.
- BATTEY, H. S. (2017). Eigen structure of a new class of covariance and inverse covariance matrices. *Bernoulli* **23**, 3166–77.
- BHATIA, R. (1997). *Matrix Analysis*. New York: Springer.
- BICKEL, P. J. & LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–604.
- BIRMAN, M. & SOLOMYAK, M. (2003). Double operator integrals in Hilbert space. *Integr. Equ. Oper. Theory* **47**, 131–68.
- BOX, G. E. P. & COX, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc. B* **26**, 211–52.
- CAI, T., LIU, W. & ZHOU, H. (2016). Estimating sparse precision matrices: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44**, 455–88.
- COX, D. R. & WERMUTH, N. (1996). *Multivariate Dependencies*. London: Chapman & Hall.
- DAVIS, C. & KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7**, 1–46.
- EFRON, B. & HASTIE, T. J. (2016). *Computer Age Statistical Inference*. Cambridge: Cambridge University Press.
- FAN, J., FENG, Y. & TONG, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *J. R. Statist. Soc. B* **74**, 745–71.
- FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179–88.
- KATO, T. (1976). *Perturbation Theory for Linear Operators*. Berlin: Springer, 2nd ed.
- MARCHENKO, V. A. & PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices [English translation]. *Math. USSR Sb.* **72**, 507–36.
- MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Hoboken, New Jersey: Wiley.
- WHITTAKER, E. T. & WATSON, G. N. (1920). *A Course of Modern Analysis*. Cambridge: Cambridge University Press, 3rd ed.

[Received on 18 October 2017. Editorial decision on 15 November 2018]