Bernoulli **23**(4B), 2017, 3166–3177
DOI: 10.3150/16-BEJ840

Eigen structure of a new class of covariance and inverse covariance matrices

HEATHER BATTEY

Department of Mathematics, Imperial College London, 180 Queen's Gate, South Kensington, London SW7 2AZ, United Kingdom; ORFE, Princeton University, 98 Charlton Street, Princeton, NJ 08540, USA.
E-mail: h.battey@imperial.ac.uk; hbattey@princeton.edu

There is a one to one mapping between a p dimensional strictly positive definite covariance matrix Σ and its matrix logarithm L. We exploit this relationship to study the structure induced on Σ through a sparsity constraint on L. Consider L as a random matrix generated through a basis expansion, with the support of the basis coefficients taken as a simple random sample of size $s = s^*$ from the index set $[p(p+1)/2] = \{1, \dots, p(p+1)/2\}$. We find that the expected number of non-unit eigenvalues of Σ, denoted $\mathbb{E}[|\mathcal{A}|]$, is approximated with near perfect accuracy by the solution of the equation

$$\frac{4p + p(p-1)}{2(p+1)}\left[\log\left(\frac{p}{p-d}\right) - \frac{d}{2p(p-d)}\right] - s^* = 0.$$

Furthermore, the corresponding eigenvectors are shown to possess only $p - |\mathcal{A}^c|$ non-zero entries. We use this result to elucidate the precise structure induced on Σ and Σ^{-1}. We demonstrate that a positive definite symmetric matrix whose matrix logarithm is sparse is significantly less sparse in the original domain. This finding has important implications in high dimensional statistics where it is important to exploit structure in order to construct consistent estimators in non-trivial norms. An estimator exploiting the structure of the proposed class is presented.

Keywords: covariance matrix; matrix logarithm; precision matrix; spectral theory

1. Introduction

In many scientific disciplines, it is natural to think of observations as n i.i.d. realisations of a p dimensional random variable $\mathbf{V}$, where $p \gg n$. This scenario is especially common in cross-sectional medical studies. For instance, in a retrospective Genome Wide Association Study (GWAS) it is natural to suppose the n patients have independent and identically distributed genome sequences $\mathbf{V}_1, \dots, \mathbf{V}_n$. Letting u denote the genomic locus, the uth element of $\mathbf{V}_i$ is the minor allele count or copy number for patient i at site u.

Numerous procedures from classical multivariate analysis (see, e.g., Anderson [1]) rely on an estimate of the $p \times p$ covariance matrix of $\mathbf{V}$, defined as $\Sigma = \mathbb{E}((\mathbf{V} - \mathbb{E}\mathbf{V})(\mathbf{V} - \mathbb{E}\mathbf{V})^T)$. So as not to cloud the presentation, we henceforth assume $\mathbb{E}\mathbf{V} = \mathbf{0}$. Despite receiving considerable attention, the problem of large covariance estimation is a persistent obstacle in numerous applied works (e.g., Cribben *et al.* [3], Mathew *et al.* [8]). It is well understood that when the dimension p is larger than the sample size n, it is impossible to construct a consistent estimator of Σ (in any non-trivial matrix norm) on the basis of the i.i.d. sample $\mathbf{V}_1, \dots, \mathbf{V}_n$ without exploiting assumed structure. Indeed, noise accumulation that results from naïvely extending low dimensional

1350-7265 © 2017 ISI/BS

procedures to high dimensional problems frequently results in classifiers that are no better than random guessing (Fan and Fan [4]) and "optimal" portfolios that are no better than a naïvely diversified one (Yuan [10]).

Just as in other branches of high dimensional estimation, suitable structural assumptions allow consistency to be restored. For instance, the assumption that $\Sigma = (\sigma_{uv})$ belongs to the class,

$$\mathcal{U}\big(q, c_0(p), M\big) = \left\{ \Sigma : \sigma_{uu} \leq M, \sum_{v=1}^{p} |\sigma_{uv}|^q \leq c_0(p) \text{ for all } u \right\} \qquad \text{for } 0 \leq q \leq 1, \tag{1.1}$$

justifies the thresholding procedure of Bickel and Levina [2], which simply sets any element of the sample covariance matrix $S_n = n^{-1} \sum_{i=1}^{n} \mathbf{V}_i \mathbf{V}_i^T$ to zero whose absolute value is below some prespecified threshold. With reference to the GWAS example above, although the sites are related by chromosomal distance, the sparsity model of equation (1.1) ignores this relationship. Instead, the model is invariant with respect to permutation of sites. This invariance may be a good thing or a bad thing, but is not to be taken for granted. Provided the threshold is chosen appropriately, the thresholding estimator is consistent in operator norm under the model in equation (1.1), guaranteeing consistency of principal components. Besides the potential implausibility of the structural assumptions imposed by (1.1), thresholding sometimes yields singular covariance estimates.

With the aim of further broadening the range of structures one can fruitfully impose on Σ and Σ^{-1}, we consider the implication on Σ and Σ^{-1} of imposing sparsity in the matrix logarithm domain. The matrix logarithm, L, of the $p \times p$ covariance matrix, Σ, is defined by $\Sigma = \exp\{L\} = \sum_{k=0}^{\infty} \frac{1}{k!} L^k$. A convenient observation is that the precision matrix, Σ^{-1}, satisfies $\Sigma^{-1} = \exp\{-L\}$. Thus any structure imposed on L induces the same structure on Σ^{-1} as it induces on Σ. In concurrent work, we explore the open problem of exploiting various forms of sparsity in the matrix logarithm domain in order define automatically positive definite estimators $\widehat{\Sigma} = \exp\{\widehat{L}\}$ and $\widehat{\Sigma}^{-1} = \exp\{-\widehat{L}\}$ for Σ and Σ^{-1} on the basis of an estimator $\widehat{L}$ of L. In this paper, we impose sparsity on L through sparsity of the coefficient vector of an expansion of L in the natural symmetrised indicator basis.

Since Σ is a positive definite symmetric matrix, L exists and is unique for Σ (Lemma 1.1). Moreover, L is of the form $L = \Xi(\log \Lambda)\Xi^T$, where $\log \Lambda = \operatorname{diag}\{\log \lambda_1, \ldots, \log \lambda_p\}$, with $\lambda_1 \geq \cdots \geq \lambda_p$ the ordered eigenvalues of Σ, and Ξ is the matrix of corresponding orthonormal eigenvectors. By its existence and uniqueness, the matrix logarithm defines a bijection between the cone of $p \times p$ symmetric positive definite matrices to which Σ belongs, and the vector space of $p \times p$ symmetric matrices,

$$\mathcal{V}(p, \mathbb{R}) := \big\{ S \in \mathcal{M}_p(\mathbb{R}) : S = S^T \big\}, \tag{1.2}$$

to which L belongs. In equation (1.2), $\mathcal{M}_p(\mathbb{R})$ is the space of $p \times p$ matrices with elements in $\mathbb{R}$. In the remainder of this paper, we use $[p]$ to denote the set of indices $\{1, \ldots, p\}$ and $|\mathcal{S}|$ to denote the cardinality of a set $\mathcal{S}$.

Lemma 1.1. *Let A be any positive definite symmetric matrix with elements in $\mathbb{R}$. Then the matrix logarithm of A exists and is unique. Moreover, it is of the form $L = \Xi(\log \Lambda)\Xi^T$, where* $\log \Lambda = \operatorname{diag}\{\log \lambda_1, \ldots, \log \lambda_p\}$, *with $\lambda_1 \geq \cdots \geq \lambda_p$ the ordered eigenvalues of A, and Ξ the matrix of corresponding orthonormal eigenvectors.*

In contrast to cones, which need not possess a cone basis, vector spaces always possess a linear basis. Indeed, there is a natural indicator basis for the vector space $\mathbb{R}^{p^2}$ of general $p \times p$ matrices, and a symmetrised indicator basis for $\mathcal{V}(p, \mathbb{R})$. This basis consists of two parts and is written $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$, where $\mathcal{B}_1$ consists of p diagonal indicator components of the form

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

for $\mathbf{e}_1, \dots, \mathbf{e}_p$ the canonical basis vectors for $\mathbb{R}^p$, and $\mathcal{B}_2$ consists of $p(p-1)/2$ symmetric non-diagonal indicator components of the form

$$\mathcal{B}_2 = \{B : B = \mathbf{e}_j \mathbf{e}_k^T + \mathbf{e}_k \mathbf{e}_j^T, j, k \in [p], j \neq k\}.$$

$\mathcal{B}$ is such that $\mathcal{V}(p, \mathbb{R}) = \text{span}_{\mathbb{R}}\{B_1, \dots, B_{p(p+1)/2}\}$ and consists of linearly independent elements of $\mathcal{V}(p, \mathbb{R})$, thus satisfying the definition of a basis.

1.1. Problem statement and notation

We consider the implication of sparsity of $\boldsymbol{\alpha}$ in the basis expansion $L(\boldsymbol{\alpha}) = \sum_{m=1}^{|\mathcal{B}|} \alpha_m B_m$ on the ordered eigenvalues $\lambda_1(\boldsymbol{\alpha}), \dots, \lambda_p(\boldsymbol{\alpha})$ of $\Sigma(\boldsymbol{\alpha}) = \exp\{L(\boldsymbol{\alpha})\}$ and corresponding eigenvectors $\boldsymbol{\xi}_1(\boldsymbol{\alpha}), \dots, \boldsymbol{\xi}_p(\boldsymbol{\alpha})$. The ordering of eigenvalues by size is simply by convention and is unnecessary apart from its role in Figures 1 and 2. We consider $\boldsymbol{\alpha}$ satisfying $\|\boldsymbol{\alpha}\|_0 = s^*$ for $\|\boldsymbol{\alpha}\|_0 = \sum_{m=1}^{|\mathcal{B}|} \mathbb{1}\{\alpha_m \neq 0\}$ and introduce the set

$$\mathcal{S} = \mathcal{S}\big(s^*(\boldsymbol{\alpha})\big) = \big\{m \in \big[p(p+1)/2\big] : \alpha_m \neq 0, \|\boldsymbol{\alpha}\|_0 = s^*\big\}, \tag{1.3}$$

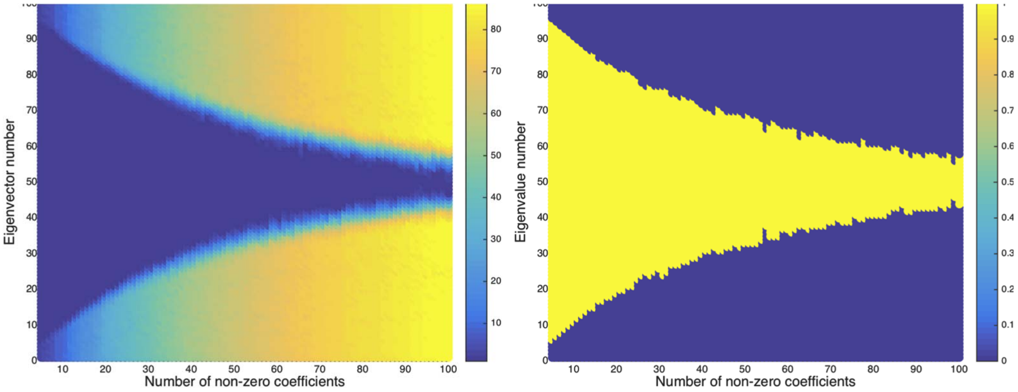
which of course satisfies $|\mathcal{S}(s^*(\boldsymbol{\alpha}))| = s^*$. In Section 2.2, we consider the support of $\boldsymbol{\alpha}$ being drawn randomly from the index set $[p(p+1)/2]$. In that case $\mathcal{S}(s^*(\boldsymbol{\alpha}))$ is a random set.

The following additional notation is used throughout. Let $\Lambda(\boldsymbol{\alpha}) = \text{diag}\{\lambda_1(\boldsymbol{\alpha}), \dots, \lambda_p(\boldsymbol{\alpha})\}$ denote the diagonal matrix of ordered (from largest to smallest) eigenvalues of $\Sigma(\boldsymbol{\alpha}) = \exp\{L(\boldsymbol{\alpha})\}$ and let $\Xi(\boldsymbol{\alpha})$ denote the corresponding matrix of eigenvectors. To ease the notational burden, we drop the reference to $\boldsymbol{\alpha}$ whenever it is unnecessary to be explicit. Then $\Sigma = \Xi \Lambda \Xi^T$ and $L = \Xi \Delta \Xi^T$, where $\Delta = \text{diag}\{\delta_1, \dots, \delta_p\}$ and $\delta_j = \log(\lambda_j)$. Hence, the unit eigenvalues of Σ correspond to the zero eigenvalues of L. Introduce the sets $\mathcal{A} := \{j \in [p] : \lambda_j \neq 1\}$, $\underline{\mathcal{A}} := \{j \in [p] : \lambda_j < 1\}$ and $\overline{\mathcal{A}} := \{j \in [p] : \lambda_j > 1\}$, then $|\mathcal{A}|$ is the number of non-unit eigenvalues of Σ and $\Xi_{\mathcal{A}}$ denotes the restriction by columns of Ξ to $\mathcal{A}$. Thus,

$$\Sigma = \Xi_{\mathcal{A}} \Lambda_{\mathcal{A}} \Xi_{\mathcal{A}}^T + \Xi_{\mathcal{A}^c} \Xi_{\mathcal{A}^c}^T = \sum_{j \in \overline{\mathcal{A}}} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T + \sum_{j \in \mathcal{A}^c \cup \underline{\mathcal{A}}} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T. \tag{1.4}$$

A special case of the second decomposition is the spiked eigenvalue model of Johnstone [6].

To set the scene for the theoretical results appearing in Section 2, Figure 1 illustrates the existence of a strong relationship between $\|\boldsymbol{\alpha}\|_0$, $\lambda_1(\boldsymbol{\alpha}), \dots, \lambda_p(\boldsymbol{\alpha})$ and $\|\boldsymbol{\xi}_1(\boldsymbol{\alpha})\|_0, \dots, \|\boldsymbol{\xi}_p(\boldsymbol{\alpha})\|_0$. In particular, fixing $p = 100$, Figure 1(A) plots the average (over 100 Monte Carlo experiments) of $\|\boldsymbol{\xi}_j(\boldsymbol{\alpha})\|_0$ as a function of $j \in [p]$ and $s^* = \|\boldsymbol{\alpha}\|_0$. Figure 1(B) plots the average of $\mathbb{1}\{\lambda_j(\boldsymbol{\alpha}) = 1\}$ as a function of $j \in [p]$ and $s^* = \|\boldsymbol{\alpha}\|_0$.



(A) (B)

Figure 1. $p = 100$ and sparsity of $\boldsymbol{\alpha}$ is $s^* \in [100]$. Panel (A) averaged (over 100 MC replications) number of non-zero eigenvector entries as a function of s^* and the corresponding eigenvalue number (ordered from largest to smallest). Panel (B) average of $\mathbb{1}\{\lambda_j(\boldsymbol{\alpha}) = 1\}$ as a function of eigenvalue number $j = 1, \dots, 100$ and s^*.

Figure 1 shows that, at high degrees of sparsity, a significant number of central eigenvectors of $\Sigma(\boldsymbol{\alpha})$ (those corresponding to unit eigenvalues) consist of a single non-zero entry and that the number of such eigenpairs is decreasing as a function of $\|\boldsymbol{\alpha}\|_0$. Moreover, for those eigenvectors possessing multiple non-zero entries, the number of non-zero entries is increasing with $\|\boldsymbol{\alpha}\|_0$. We explain the relationship theoretically in Section 2.

2. Theoretical results

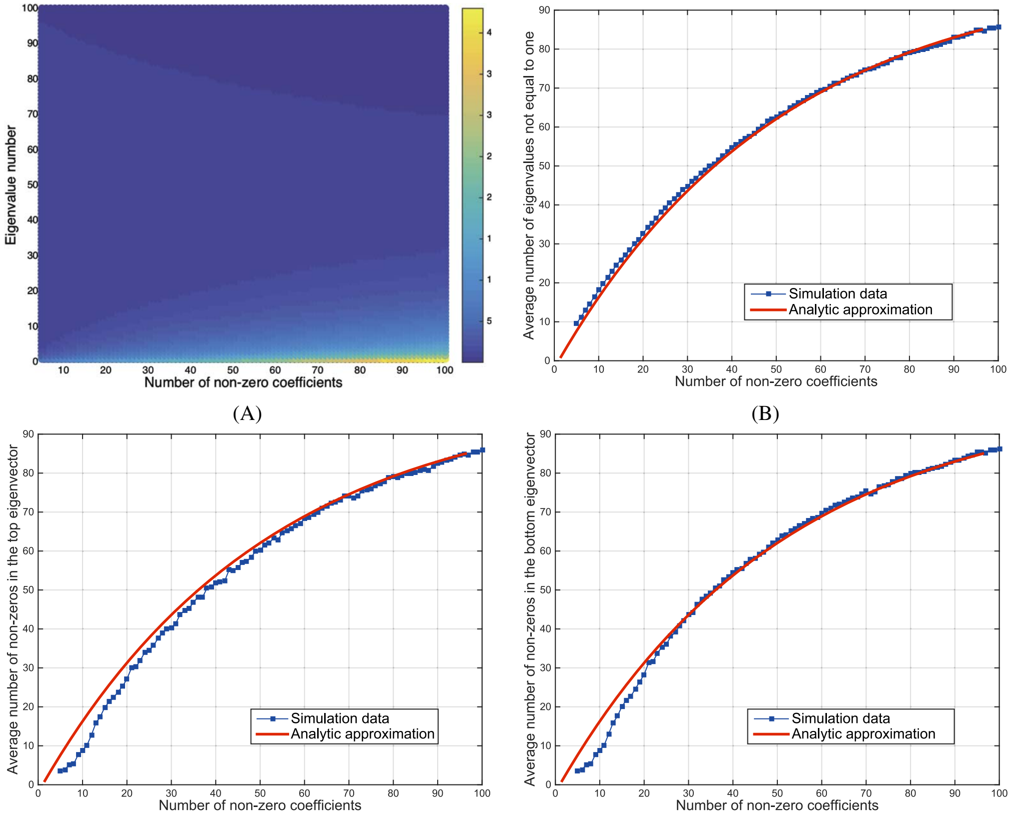
Our theoretical results of this section relate the structure in the eigendecomposition of $\Sigma(\boldsymbol{\alpha})$ to the sparsity structure of L, which we define through sparsity of $\boldsymbol{\alpha}$ in the basis expansion $L(\boldsymbol{\alpha}) = \sum_{m=1}^{|\mathcal{B}|} \alpha_m B_m$. The first results of this section hold independently of the random sampling of the support of $\boldsymbol{\alpha}$. In Section 2.2, we explore the expected number of non-unit eigenvalues of $\Sigma(\boldsymbol{\alpha}) = \exp\{L(\boldsymbol{\alpha})\}$ when the support of $\boldsymbol{\alpha}$ is a simple random sample of size $s = s^*$ from the index set $[p(p+1)/2]$.

2.1. Some preliminary deterministic results

The first result establishes that, for every $\boldsymbol{\alpha}$, $|\mathcal{A}|$ is equal to the expected number of distinct column vectors of $\mathcal{B}_{\mathcal{S}} := \{B_m \in \mathcal{B} : m \in \mathcal{S}(s^*(\boldsymbol{\alpha}))\}$. We denote the collection of non-zero column vectors of $\mathcal{B}_{\mathcal{S}}$ as

$$\mathcal{L}\big(s^*(\boldsymbol{\alpha})\big) := \big\{\mathbf{b}_j^{(m)} : j \in [p], \mathbf{b}_j^{(m)} \neq \mathbf{0}, m \in \mathcal{S}\big(s^*(\boldsymbol{\alpha})\big)\big\}, \tag{2.1}$$

where $\mathbf{b}_j^{(m)}$ for $m \in \mathcal{S}$ is the jth column of the mth element of $\mathcal{B}_{\mathcal{S}} \subset \mathcal{B}$.



(A)

(B)

(C)

(D)

Figure 2. $p = 100$ and sparsity of $\boldsymbol{\alpha}$ is $s^* \in [100]$. Panel (A): averaged (over 100 MC replications) eigenvalue of $\exp\{L\}$ as a function of eigenvalue number and s^*. Panel (B): average value of $|\mathcal{A}|$ as a function of s^*. The red line is the theoretical value based on equation (2.4). Panels (C) and (D): average value of $\|\boldsymbol{\xi}_1\|_0$ and $\|\boldsymbol{\xi}_p\|_0$ as a function of s^*.

Lemma 2.1. *For any* $\boldsymbol{\alpha} \in \mathbb{R}^{p(p+1)/2}$, $|\mathcal{A}| = D^*(\boldsymbol{\alpha})$, *where* $D^*(\boldsymbol{\alpha})$ *is the number of distinct elements of the collection* $\mathcal{L}(s^*(\boldsymbol{\alpha}))$ *and* $\mathcal{A} = \{j \in [p] : \lambda_j \neq 1\}$.

It is immediately clear that $D^*(\boldsymbol{\alpha}) \leq |\mathcal{L}(s^*(\boldsymbol{\alpha}))| \leq 2s^*$ because each element of $\mathcal{B}_{\mathcal{S}}$ contains at most 2 non-zero columns. However, the exact value of $D^*(\boldsymbol{\alpha})$ depends on the support of $\boldsymbol{\alpha}$. In Section 2.2, we consider drawing the support of $\boldsymbol{\alpha}$ at random from the index set $[p(p+1)/2]$. Under simple random sampling, $\mathcal{B}_{\mathcal{S}}$ has a non-zero probability of containing basis matrices in $\mathcal{B}_1$. More importantly, the probability that columns are repeated is increasing in s^* implying that the bound $2s^*$ becomes more conservative as s^* increases. This observation is visually apparent in Figure 2, where we also display an accurate analytic approximation, derived in Section 2.2.

The next lemma shows that the behaviour of the eigenvectors is also explained by $D^*(\boldsymbol{\alpha})$.

Lemma 2.2. *For any $\boldsymbol{\alpha} \in \mathbb{R}^{p(p+1)/2}$, there exists an orthonormal set of eigenvectors $\boldsymbol{\xi}_1(\boldsymbol{\alpha}), \dots, \boldsymbol{\xi}_p(\boldsymbol{\alpha})$ of $\Sigma(\boldsymbol{\alpha})$ such that for any $j \in \mathcal{A}^c$, $\boldsymbol{\xi}_j$ is of the form $\boldsymbol{\xi}_j = \mathbf{e}_{v(j)} \neq \boldsymbol{\xi}_\ell$ for $\mathbf{e}_{v(j)} \notin \mathcal{L}(s^*(\boldsymbol{\alpha}))$ and $\ell \in \mathcal{A}^c$. Additionally, for all $j \in \mathcal{A}$, $\|\boldsymbol{\xi}_j\|_0 = p - |\mathcal{A}^c|$.*

Remark 2.3. The matrix of eigenvectors, $\Xi = [\Xi_{\mathcal{A}}, \Xi_{\mathcal{A}^c}]$, is unique up to permutations of the columns of $\Xi_{\mathcal{A}^c}$.

Lemmata 2.1 and 2.2 together give rise to the covariance model described in Corollary 2.4.

Corollary 2.4. *For any $\boldsymbol{\alpha}$, $\Sigma(\boldsymbol{\alpha})$ is of the form*

$$\Sigma = \sum_{j \in \overline{\mathcal{A}}} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T + \sum_{j \in \mathcal{A}^c} \mathbf{e}_{v(j)} \mathbf{e}_{v(j)}^T + \sum_{j \in \underline{\mathcal{A}}} \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T.$$

Above, $|\mathcal{A}| = |\underline{\mathcal{A}} \cup \overline{\mathcal{A}}| = D^*(\boldsymbol{\alpha})$ for $D^*(\boldsymbol{\alpha})$ is defined in Lemma 2.1 and where, for all $j \in \mathcal{A}$, $\|\boldsymbol{\xi}_j\|_0 = p - |\mathcal{A}^c|$.

2.2. An analytic approximation to the expected number of non unit eigenvalues

We use the deterministic results from the previous section to obtain an expression (in terms of p and s^*) for the expected value of $|\mathcal{A}| = |\mathcal{A}(\boldsymbol{\alpha})|$ when the support of $\boldsymbol{\alpha}$ is a simple random sample of size $s = s^*$ from the index set $[p(p+1)/2]$. Such a formula is unavailable in closed form. However, based on Lemma 2.1 we provide a very close analytic approximation in terms of p and s^* (the red curve in panels (B), (C) and (D) of Figure 2).

Consider an approximation to the expected number of random draws, N_d, from $\mathcal{B}$ required to obtain d distinct elements of $\mathcal{L}(N_d(\boldsymbol{\alpha}))$. Replacing the expected number of draws by s^* and solving for d yields, in view of Lemma 2.1, an approximation to the expected number of non unit eigenvalues of $\Sigma(\boldsymbol{\alpha})$ as a function of p and s^*.

Consider $\mathcal{L}(s^*) := \mathcal{L}(s^*(\boldsymbol{\alpha}))$ of equation (2.1) as an element of the nested sequence of random sets $\mathcal{L}(1) \subseteq \mathcal{L}(2) \subseteq \cdots$. Let $X_1 = 1$ denote the number of draws from $\mathcal{B}$ required to obtain the first new non-zero column vector in $\mathcal{L}(1)$. For $i \geq 2$, let X_i be the number of draws required to obtain the ith new column vector of $\mathcal{L}(\sum_{j=1}^{i-1} X_j)$ after the $(i-1)$th new column has been obtained. Then the expected number of draws, N_d such that $\mathcal{L}(N_d)$ contains d distinct columns satisfies

$$\begin{aligned} \mathbb{E}[N_d] &\approx \sum_{i=1}^{d} \left[\frac{2p}{p(p+1)} \mathbb{E}[X_i | \mathcal{B}_1] + \frac{p-1}{p+1} \mathbb{E}[X_i | \mathcal{B}_2] \right]. \\ &= \frac{4p + p(p-1)}{2(p+1)} \sum_{i=1}^{d} \frac{1}{p-(i-1)}, \end{aligned} \tag{2.2}$$

where the last line is obtained by noting that X_i is a geometric random variable with parameter $(p-(i-1))/p$ regardless of the distribution from which the nonzero values of $\alpha_1, \dots, \alpha_{|\mathcal{B}|}$ are drawn. The left-hand side of equation (2.2) is only approximately equal to the right-hand side because we approximate sampling without replacement from $\mathcal{B}$ by sampling with replacement from $\mathcal{B}$. This approximation is accurate when $p(p+1)/2$ is large relative to s^*. An approximate expression for the expected number of distinct columns of $\mathcal{L}(s^*(\boldsymbol{\alpha}))$ is thus obtained by setting the left-hand side of equation (2.2) to s^* and solving for d, where d appears in the expression for $\sum_{i=1}^{d} \frac{1}{p-(i-1)}$. This expression is provided in equation (2.3). Changing variables to $j = p - (i-1)$, we have

$$\begin{aligned}
\mathbb{E}[N_d] &\approx \frac{4p + p(p-1)}{2(p+1)} \sum_{j=p-(d-1)}^{p} \frac{1}{j} = \frac{4p + p(p-1)}{2(p+1)} \left[\sum_{j=1}^{p} \frac{1}{j} - \sum_{j=1}^{p-d} \frac{1}{j} \right] \\
&= \frac{4p + p(p-1)}{2(p+1)} \left[\log p + \gamma + \varepsilon_p - \left(\log(p-d) + \gamma + \varepsilon_{p-d} \right) \right] \qquad (2.3) \\
&= \frac{4p + p(p-1)}{2(p+1)} \left[\log\left(\frac{p}{p-d} \right) + (\varepsilon_p - \varepsilon_{p-d}) \right],
\end{aligned}$$

where γ is the Euler–Mascheroni constant and $\varepsilon_p \approx 1/2p$ and $\varepsilon_{p-d} \approx 1/2(p-d)$. It follows that the expected number of distinct elements of $\mathcal{L}(s^*(\boldsymbol{\alpha}))$ from Lemma 2.1 is well approximated by the solution, d^*, of

$$\frac{4p + p(p-1)}{2(p+1)} \left[\log\left(\frac{p}{p-d} \right) - \frac{d}{2p(p-d)} \right] = s^*,$$

that is,

$$d^* = \operatorname{root}\left\{ \frac{4p + p(p-1)}{2(p+1)} \left[\log\left(\frac{p}{p-d} \right) - \frac{d}{2p(p-d)} \right] - s^* \right\}. \qquad (2.4)$$

We plot this solution for $p = 100$ in Figure 2, observing that the analytic approximation derived above coincides almost perfectly with the numerical results.

2.3. Implications and discussion

Lemmata 2.1 and 2.2 together with equation (2.4) straightforwardly yield the following result, which is independent of distributional assumptions on $\boldsymbol{\alpha}$ and only relies on the simple random sampling of the support of $\boldsymbol{\alpha}$ from the index set $[p(p+1)/2]$.

Theorem 2.5. *A random positive definite symmetric matrix $\Sigma(\boldsymbol{\alpha})$ is logarithmically sparse in the sense that*

$$\Sigma(\boldsymbol{\alpha}) = \exp\{L(\boldsymbol{\alpha})\} \qquad \text{with } L(\boldsymbol{\alpha}) = \sum_{m=1}^{M} \alpha_m B_m \text{ and } \|\boldsymbol{\alpha}\|_0 = s^*$$

if and only if $\Sigma(\boldsymbol{\alpha})$ is of the form $\Sigma(\boldsymbol{\alpha}) = PKP^{-1}$ with P a permutation matrix and K a block diagonal matrix with blocks K_1 and I_{p-D^*}, $\mathbb{E}(D^*) = d^*$ (cf. equation (2.4)).

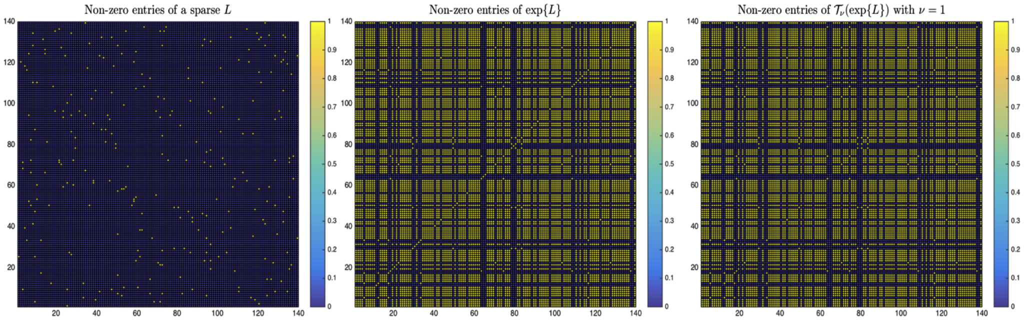


Figure 3. (Left) Non-zero elements of a sparse L, where sparsity of L is measured by the number of non-zero coefficients of its basis expansion; (centre) non-zero elements of $\Sigma = \exp\{L\}$, where L is that depicted in the left panel; (right) non-zero elements of Σ threshold at $\nu = 1$.

One advantage of exploring sparsity in the matrix logarithm domain is that a sparse L potentially gives rise to a $\Sigma = \exp\{L\}$ and $\Sigma^{-1} = \exp\{-L\}$ which are significantly less sparse. Figure 3 illustrates this fact.

In terms of the random vector $\mathbf{V}$ itself, sparsity of $\boldsymbol{\alpha}$ implies that $\mathbf{V}$ can be decomposed into two subsets of variables, $\mathbf{V}_1$ and $\mathbf{V}_2$ such that $\mathbf{V}_1$ has covariance structure K_1, whilst the elements of $\mathbf{V}_2$ are completely uncorrelated with each other and with the elements of $\mathbf{V}_1$. This naturally raises the question of whether matrix logarithmic sparsity is more or less plausible than model (1.1). With $q = 0$, model (1.1) implies all variables are uncorrelated with all but $c_0(p)$ of the others, where $c_0(p)$ must be such that $(c_0(p) \log p)/n \to 0$ for the model in class (1.1) to be statistically useful. By contrast, a matrix logarithmically sparse model assumes that a large group of variables are arbitrarily correlated with others in the same group but completely uncorrelated with those in another group, which in turn are uncorrelated with each other. There are undoubtedly examples for sparsity on every scale, perhaps having removed common factors as in Fan *et al.* [5]. The difficulty in assessing the plausibility of these models a priori highlights the need for further work in the area. In the context of bandable covariance matrices (Wu and Pourahmadi [9]), Zou and Li [11] develop an information criterion for selecting the tuning parameters of the banding estimator. In principal at least, it should be possible to develop information criteria for selecting between different classes of covariance model.

The structural assumptions on L are naturally exploited through a penalised regression-based estimator of L. More specifically, letting $\widehat{L}^P$ denote a elementwise consistent pilot estimator for L, letting $\|\cdot\|_F$ denote the Frobenius norm, and letting $\mathcal{P}(\eta) = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_1 \leq \eta\}$, an estimator of L is constructed as $\widehat{L} = \sum_{m=1}^{|\mathcal{B}|} \widehat{\alpha}_m B_m$ where

$$\widehat{\boldsymbol{\alpha}} = \underset{\alpha \in \mathbb{R}^{|\mathcal{B}|} \cap \mathcal{P}(\eta)}{\operatorname{argmin}} \left\{ \left\| \widehat{L}^P - \sum_{m=1}^{|\mathcal{B}|} \alpha_m B_m \right\|_F^2 \right\}.$$

By performing the minimisation over $\mathbb{R}^{|\mathcal{B}|} \cap \mathcal{P}(\eta)$, we exploit the assumed sparsity in the basis coefficient vector $\boldsymbol{\alpha}$. In on-going work, we are investigating this new estimator and will report the result elsewhere.

3. Proofs

Proof of Lemma 1.1. We first prove the existence result. Since A is symmetric, it is orthogonally diagonalisable by the spectral theorem, with an orthonormal eigendecomposition $A = \Xi\Lambda\Xi^T$ that is unique up to permutations of the columns of Ξ corresponding to the repeated eigenvalues in $\Lambda = \mathrm{diag}\{\lambda_1, \dots, \lambda_p\}$. Let $L = \Xi(\log\Lambda)\Xi^T$, where $\log\Lambda = \mathrm{diag}\{\log\lambda_1, \dots, \log\lambda_p\}$. Then L is a matrix logarithm of A because

$$\begin{aligned}\exp\{L\} &= \sum_{k=0}^{\infty}\frac{1}{k!}L^k = \Xi\left(\sum_{k=0}^{\infty}\frac{1}{k!}(\log\Lambda)^k\right)\Xi^T \\ &= \Xi\big(\mathrm{diag}\big\{\exp\{\log\lambda_1\}, \dots, \exp\{\log\lambda_p\}\big\}\big)\Xi^T = A.\end{aligned}$$

To prove uniqueness, let L' be another matrix satisfying $A = \exp\{L'\} = \sum_{k=0}^{\infty}\frac{1}{k!}(L')^k$. Since A is symmetric, so is L', thus L' is orthogonally diagonalisable with orthonormal eigendecomposition $L' = ODO^T$, which is unique up to permutations of the columns of O corresponding to repeated eigenvalues in $D = \mathrm{diag}\{d_1, \dots, d_p\}$. We have

$$A = \exp\big\{L'\big\} = O\left(\sum_{k=0}^{\infty}\frac{1}{k!}D^k\right)O^T = O\big(\mathrm{diag}\big\{\exp\{d_1\}, \dots, \exp\{d_p\}\big\}\big)O^T.$$

By the uniqueness of Ξ and O up to permutations of the columns corresponding to repeated eigenvalues, and since the exponential function is an isomorphism, we know that $d_j = \log\lambda_j$ and $\Xi = O$, thus $L = L'$. □

Proof of Lemma 2.1. The first part of the proof is to show that the number of non-zero eigenvalues of L is equal to the dimension of the image of L. To this end, recall the definitions of the kernel (null space) and the image (column space) of L:

$$\mathrm{Ker}(L) := \big\{\mathbf{x} \in \mathbb{R}^p : L\mathbf{x} = \mathbf{0}\big\}, \qquad \mathrm{Im}(L) := \big\{\mathbf{y} \in \mathbb{R}^p : \exists\mathbf{x} \in \mathbb{R}^p \text{ for which } L\mathbf{x} = \mathbf{y}\big\}. \tag{3.1}$$

We first demonstrate that $\{\boldsymbol{\xi}_j : j \in \mathcal{A}^c\}$ is a basis for $\mathrm{Ker}(L)$. Since eigenvectors are linearly independent by their orthogonality, this simply amounts to showing that $\mathrm{span}_{\mathbb{R}}\{\boldsymbol{\xi}_j : j \in \mathcal{A}^c\} = \mathrm{Ker}(L)$, i.e. $\mathrm{span}_{\mathbb{R}}\{\boldsymbol{\xi}_j : j \in \mathcal{A}^c\} \subseteq \mathrm{Ker}(L)$ and $\mathrm{Ker}(L) \subseteq \mathrm{span}_{\mathbb{R}}\{\boldsymbol{\xi}_j : j \in \mathcal{A}^c\}$.

The first containment is trivial. To prove $\mathrm{Ker}(L) \subseteq \mathrm{span}_{\mathbb{R}}\{\boldsymbol{\xi}_j : j \in \mathcal{A}^c\}$, suppose for a contradiction that there exists $\mathbf{v} \in \mathrm{Ker}(L)$ such that $\mathbf{v} \notin \mathrm{span}\{\boldsymbol{\xi}_j : j \in \mathcal{A}^c\}$. Without loss of generality denote the columns of $\Xi_{\mathcal{A}^c}$ by $\{\widetilde{\boldsymbol{\xi}}_1, \dots, \widetilde{\boldsymbol{\xi}}_k\}$ so that the columns of $\Xi_{\mathcal{A}}$ are $\{\widetilde{\boldsymbol{\xi}}_{k+1}, \dots, \widetilde{\boldsymbol{\xi}}_p\}$ (we have used $\{\widetilde{\boldsymbol{\xi}}_j : j \in [p]\}$ to differentiate the unordered eigenvectors from the ordered ones $\{\boldsymbol{\xi}_j : j \in [p]\}$). For any vector space $V \subset \mathbb{R}^p$, any set of $k \leq p$ linearly independent vectors is either a basis for V or can be extended to a basis. Letting $\{\widetilde{\boldsymbol{\xi}}_1, \dots, \widetilde{\boldsymbol{\xi}}_k, \boldsymbol{\xi}'_{k+1}, \dots, \boldsymbol{\xi}'_p\}$ denote an extended basis, we may write

$$\mathbf{v} = \beta_1\widetilde{\boldsymbol{\xi}}_1 + \cdots + \beta_k\widetilde{\boldsymbol{\xi}}_k + \beta_{k+1}\boldsymbol{\xi}'_{k+1} + \cdots + \beta_p\boldsymbol{\xi}'_p$$

and by the fact that $\widetilde{\boldsymbol{\xi}}_1, \dots, \widetilde{\boldsymbol{\xi}}_k \in \operatorname{Ker}(L)$,

$$\begin{aligned}L\mathbf{v} &= \beta_1 L\widetilde{\boldsymbol{\xi}}_1 + \cdots + \beta_k L\widetilde{\boldsymbol{\xi}}_k + \beta_{k+1} L\boldsymbol{\xi}'_{k+1} + \cdots + \beta_p L\boldsymbol{\xi}'_p \\ &= \beta_{k+1} L\boldsymbol{\xi}'_{k+1} + \cdots + \beta_p L\boldsymbol{\xi}'_p.\end{aligned}$$

Since $\{\widetilde{\boldsymbol{\xi}}_1, \dots, \widetilde{\boldsymbol{\xi}}_k, \boldsymbol{\xi}'_{k+1}, \dots, \boldsymbol{\xi}'_p\}$ is a basis for V, $\boldsymbol{\xi}'_{k+1}, \dots, \boldsymbol{\xi}'_p$ are linearly independent of $\widetilde{\boldsymbol{\xi}}_1, \dots, \widetilde{\boldsymbol{\xi}}_k$, and by the fact that $\{\widetilde{\boldsymbol{\xi}}_1, \dots, \widetilde{\boldsymbol{\xi}}_k, \widetilde{\boldsymbol{\xi}}_{k+1}, \dots, \widetilde{\boldsymbol{\xi}}_p\}$ forms a basis for $\mathbb{R}^p$, there exist $\gamma_{k+1}^{(j)}, \dots, \gamma_p^{(j)}$ such that $\boldsymbol{\xi}'_j = \sum_{\ell=k+1}^p \gamma_\ell^{(j)} \widetilde{\boldsymbol{\xi}}_\ell$ for any $j \in \{k+1, \dots, p\}$. Hence, using the fact that $\mathbf{v} \in \operatorname{Ker}(L)$, we have

$$\begin{aligned}L\mathbf{v} &= \beta_{k+1} L\left(\sum_{\ell=k+1}^p \gamma_\ell^{(k+1)} \widetilde{\boldsymbol{\xi}}_\ell\right) + \cdots + \beta_p L\left(\sum_{\ell=k+1}^p \gamma_\ell^{(p)} \widetilde{\boldsymbol{\xi}}_\ell\right) \\ &= \beta_{k+1} \sum_{\ell=k+1}^p \gamma_\ell^{(k+1)} L\widetilde{\boldsymbol{\xi}}_\ell + \cdots + \beta_p \sum_{\ell=k+1}^p \gamma_\ell^{(p)} L\widetilde{\boldsymbol{\xi}}_\ell = \mathbf{0}.\end{aligned} \tag{3.2}$$

Equation (3.2) implies that either

$$\beta_{k+1} = \gamma_{k+1}^{(k+1)} = \cdots = \gamma_p^{(k+1)} = \cdots = \gamma_{k+1}^{(p)} = \cdots = \gamma_p^{(p)} = 0,$$

contradicting linear independence, or that $L\widetilde{\boldsymbol{\xi}}_\ell = \mathbf{0}$ for at least one $\ell \in \{k+1, \dots, p\}$, contradicting the fact that $\widetilde{\boldsymbol{\xi}}_\ell$ is a column of $\Xi_{\mathcal{A}}$.

By linear independence, $|\mathcal{A}| = \dim(\operatorname{span}_{\mathbb{R}}\{\widetilde{\xi}_{k+1}, \dots, \widetilde{\xi}_p\})$ and by our previous demonstration we have $\dim(\operatorname{span}_{\mathbb{R}}\{\widetilde{\xi}_{k+1}, \dots, \widetilde{\xi}_p\}) = \dim(\operatorname{Ker}(L))$. By the Rank Nullity Theorem (Körner [7]),

$$p = |\mathcal{A} \cup \mathcal{A}^c| = |\mathcal{A}| + |\mathcal{A}^c| = \dim(\operatorname{Ker}(L)) + \dim(\operatorname{Im}(L))$$

and therefore $|\mathcal{A}| = \dim(\operatorname{Im}(L))$. Let $\mathbf{a}_1, \dots, \mathbf{a}_p$ denote the columns of L. Then $L\mathbf{x} = x_1\mathbf{a}_1 + \cdots + x_p\mathbf{a}_p$, but through the sparsity constraint on the basis expansion of L, $L = \sum_{m\in\mathcal{S}} \alpha_m B_m$ (cf. equation (1.3)), hence letting $\mathbf{b}_j^{(m)}$ denote the jth column of B_m,

$$L\mathbf{x} = x_1\left(\sum_{m\in\mathcal{S}} \alpha_m \mathbf{b}_1^{(m)}\right) + \cdots + x_p\left(\sum_{m\in\mathcal{S}} \alpha_m \mathbf{b}_p^{(m)}\right) = \sum_{m\in\mathcal{S}} \sum_{j=1}^p (\alpha_m x_j)\mathbf{b}_j^{(m)}$$

so $\operatorname{Im}(L) = \operatorname{span}\{\mathbf{b}_j^{(m)} : m \in \mathcal{S}, j \in [p]\}$ and since all $\mathbf{b}_j^{(m)} \in \{\mathbf{0}, \mathbf{e}_1, \dots, \mathbf{e}_p\}$ the dimension of $\operatorname{Im}(L)$ is simply the number of distinct column vectors in the matrix elements $B_1^{\mathcal{S}}, \dots, B_{s^*}^{\mathcal{S}}$ of $\mathcal{B}_{\mathcal{S}}$, i.e. D^*. □

Proof of Lemma 2.2. As shown in the proof of Lemma 2.1, $\boldsymbol{\xi}_j$ is a column of $\Xi_{\mathcal{A}^c}$ if and only if $\boldsymbol{\xi}_j \in \operatorname{Ker}(L)$. Let $\mathcal{V} \subset [p]$ denote the set of indices corresponding to the nonzero row coordinates of the elements of $\mathcal{L}(s^*(\boldsymbol{\alpha}))$. Then for any $j \in \mathcal{A}^c = \{j \in [p] : \lambda_j = 1\}$,

$$L\boldsymbol{\xi}_j = \boldsymbol{\xi}_j\left(\sum_{m\in\mathcal{S}} \alpha_m \mathbf{b}_1^{(m)}\right) + \cdots + \boldsymbol{\xi}_j\left(\sum_{m\in\mathcal{S}} \alpha_m \mathbf{b}_p^{(m)}\right) = \mathbf{0}. \tag{3.3}$$

Equation (3.3) is satisfied for any $\boldsymbol{\xi}_j$ for which the only nonzero coordinate, say $v(j)$ belongs to $\mathcal{V}^c$. By Lemma 2.1, there are $|\mathcal{A}| = p - |\mathcal{A}^c|$ such possible coordinates. All other $\{\boldsymbol{\xi}_\ell : \ell \in \mathcal{A}^c\}$ are specified similarly, where orthonormality requires that any two $\boldsymbol{\xi}_j$ and $\boldsymbol{\xi}_\ell$ such that $j, \ell \in \mathcal{A}^c$ contain a single one at coordinates $v(j)$ and $v(\ell)$ respectively, where $v(j) \neq v(\ell)$ and $v(j)$ and $v(\ell)$ both belong to $\mathcal{V}^c$. We later demonstrate that a particular eigenvector solution with $\{\boldsymbol{\xi}_j : j \in \mathcal{A}^c\}$ is unique up to permutations of the columns of $\Xi_{\mathcal{A}^c}$.

By similar calculations to those appearing the proof of Lemma 2.1, $\mathrm{Im}(L) = \mathrm{span}_{\mathbb{R}}\{\boldsymbol{\xi}_j : j \in \mathcal{A}\}$, where the image (column space) of L is defined in equation (3.1), and $\dim(\mathrm{Im}(L)) = D^*(\boldsymbol{\alpha})$ as derived in the proof of Lemma 2.1. The elements of $\{\boldsymbol{\xi}_j : j \in \mathcal{A}\}$ form the columns of $\Xi_{\mathcal{A}}$, and by linear independence of these columns together with Lemma 2.1, the column space of $\Xi_{\mathcal{A}}$ is $|\mathcal{A}| = D^*(\boldsymbol{\alpha})$. Since the column space and the row space of a matrix must coincide, only $p - |\mathcal{A}^c|$ rows of $\Xi_{\mathcal{A}}$ are linearly dependent. Eigenvectors $\{\xi_j : j \in \mathcal{A}\}$ orthogonal to the $\{\xi_j : j \in \mathcal{A}^c\}$ eigenvectors constructed above can be obtained by setting the $p - |\mathcal{A}^c|$ linearly dependent rows of $\Xi_{\mathcal{A}}$ equal to zero. Since, for any symmetric matrix M, there is a unique orthonormal set of eigenvectors, up to permutations of the columns corresponding to repeated eigenvalues, we have proved that $\{\xi_j : j \in \mathcal{A}\}$ and $\{\xi_j : j \in \mathcal{A}^c\}$ possess the sparsity structure of Lemma 2.2. □

Acknowledgements

The author is grateful to Jianqing Fan for helpful discussions related to this work. She also thanks the Editor and Associate Editor for their thoughtful comments and suggestions, which greatly enhanced the clarity of the work.

References

[1] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR1990662

[2] Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[3] Cribben, I., Haraldsdottir, R., Atlas, L., Wager, T. and Lindquist, M. (2012). Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage* **61** 907–920.

[4] Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

[5] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653

[6] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[7] Körner, T.W. (2013). *Vectors*, *Pure and Applied*: *A General Introduction to Linear Algebra*. Cambridge: Cambridge Univ. Press. MR3014686

[8] Mathew, B., Holand, A.M., Koistinen, P., Léon, J. and Sillanpää, M.J. (2016). Reparametrization-based estimation of genetic parameters in multi-trait animal model using integrated nested Laplace approximation. *Theor. Appl*. *Genet*. **129** 215–225.

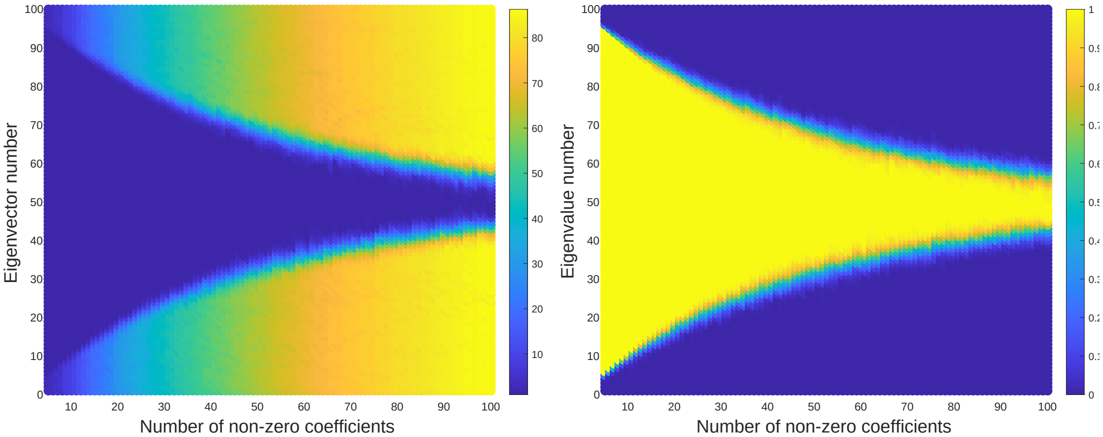
[9] Wu, W.B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. MR2024760

[10] Yuan, M. and Chen, J. (2016). Efficient Portfolio Selection in a Large Market. *J. Financ*. *Econom*. To appear.

[11] Zou, H. and Li, D. (2016). SURE information criteria for large covariance matrix estimation and their asymptotic properties. *IEEE Trans*. *Inform*. *Theory* **62** 2153–2169.

Received December 2015 and revised March 2016

Note added after publication. Figure 1 should appear as:


Eigenvector number
Number of non-zero coefficients
Eigenvalue number
Number of non-zero coefficients