Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **52**(4), 2010, 423–437 doi: 10.1111/j.1467-842X.2010.00590.x

ON ESTIMATION OF THE POPULATION SPECTRAL DISTRIBUTION FROM A HIGH-DIMENSIONAL SAMPLE COVARIANCE MATRIX

ZHIDONG BAI[1], JIAQI CHEN[1,2] AND JIANFENG YAO[2*]

National University of Singapore, Northeast Normal University and Université de Rennes 1

Summary

Sample covariance matrices play a central role in numerous popular statistical methodologies, for example principal components analysis, Kalman filtering and independent component analysis. However, modern random matrix theory indicates that, when the dimension of a random vector is not negligible with respect to the sample size, the sample covariance matrix demonstrates significant deviations from the underlying population covariance matrix. There is an urgent need to develop new estimation tools in such cases with high-dimensional data to recover the characteristics of the population covariance matrix from the observed sample covariance matrix. We propose a novel solution to this problem based on the method of moments. When the parametric dimension of the population spectrum is finite and known, we prove that the proposed estimator is strongly consistent and asymptotically Gaussian. Otherwise, we combine the first estimation method with a cross-validation procedure to select the unknown model dimension. Simulation experiments demonstrate the consistency of the proposed procedure. We also indicate possible extensions of the proposed estimator to the case where the population spectrum has a density.

Key words: eigenvalues of covariance matrices; high-dimensional statistics; Marčenko–Pastur distribution; sample covariance matrices.

1. Introduction

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sequence of independent and identically distributed (i.i.d.) zero-mean random vectors in $\mathbb{R}^p$ or $\mathbb{C}^p$, with a common population covariance matrix Σ_p. Almost all statistical theories dealing with large samples were developed through probabilistic limiting theorems of fixed dimension p and increasing sample sizes n. However, modern random matrix theory (RMT) predicts that, when the dimension of $\mathbf{x}_i$, p, is not negligible with respect to the sample size n, the sample covariance matrix,

$$\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top,$$

does not approach Σ_p. Therefore, classical statistical procedures based on an approximation of Σ_p by $\mathbf{S}_n$ become inconsistent or very inefficient in situations with high-dimensional data. There is thus a great need to develop new statistical tools for high-dimensional data analysis.

*Author to whom correspondence should be addressed.
[1]KLASMOE and School of Mathematics and Statistics, Northeast Normal University, 5268 People's Road, 130024 Changchun, China.
[2]IRMAR, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France.
e-mail: jian-feng.yao@univ-rennes1.fr

Acknowledgments. The authors wish to thank the Chinese National Science Foundation, Northeast Normal University (China) and Région Bretagne (France) for their support of this research.

As a historical example, Dempster (1958, 1960) proposed a so-called *non-exact test* to correct the deficiency of the Hotelling T^2-test.

From among these challenging statistical problems, we address one aiming to recover the characteristics of the population covariance matrix Σ_p from the sample covariance matrix $\mathbf{S}_n$. This problem is of central importance in popular statistical methodologies such as principal components analysis (Johnstone 2001), Kalman filtering and independent component analysis, which all rely on efficient estimation of some population covariance matrices.

Let $(s_j)_{1\le j\le p}$ be the p eigenvalues of Σ_p. We are particularly interested in the following spectral distribution (SD) H_p of Σ_p, that is, the distribution

$$H_p = \frac{1}{p}\sum_{j=1}^{p} \delta_{s_j},$$

where δ_b denotes the Dirac point measure at b.

Furthermore, let $(\lambda_j)_{1\le j\le p}$ be the so-called sample eigenvalues, that is, the eigenvalues of $\mathbf{S}_n$ (assuming $p \le n$). Analogously, the empirical SD (ESD) of $\mathbf{S}_n$ is the random measure

$$F_n = \frac{1}{p}\sum_{j=1}^{p} \delta_{\lambda_j}.$$

As mentioned above, for high-dimensional data, $\mathbf{S}_n$ does not approach Σ_p, and consequently F_n deviates from H_p. While considering $n, p \to \infty$, it is natural to assume that H_p weakly converges to a limiting distribution H. We refer to this limiting SD H as the *population spectral distribution* (PSD) of the observation model. For instance, an important situation we will consider is when the PSD H has a finite support $\{a_1, \ldots, a_k\} \subset \mathbb{R}_+$:

$$H = \sum_{j=1}^{k} t_j \delta_{a_j}, \tag{1}$$

where $t_j > 0$ and $t_1 + \cdots + t_k = 1$.

The following important question is then raised: how do we recover the PSD H from the sample covariance matrix $\mathbf{S}_n$? Recently, El Karoui (2008) proposed a variational and nonparametric approach to this problem, based on an appropriate distance function using (3) below and a large dictionary made with base density functions and Dirac point masses. The estimator proposed by El Karoui (2008) is proved consistent in a nonparametric estimation sense, assuming that both the dictionary size and the number of observations n tend to infinity. However, no result on the convergence rate of the estimator, for example a central limit theorem, is given.

In another important work, Raj Rao *et al.* (2008) proposed using a suitable set of empirical moments

$$\hat{\alpha}_j = \frac{1}{p}\operatorname{tr}\mathbf{S}_n^j = \frac{1}{p}\sum_{\ell=1}^{p} \lambda_\ell^j, \quad j = 1, \ldots, q.$$

More precisely, when $n \to \infty$, the mean and variance parameters, respectively $\mathbf{m}_\theta$ and $\mathbf{Q}_\theta$, of the Gaussian limiting distributions of the sample moments $(\hat{\alpha}_j)$ are known functions of the unknown parameters θ of H. Their estimator $\hat{\theta}$ is obtained by maximizing the Gaussian

likelihood, letting $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_j)_{1 \le j \le q}$,

$$\exp\left[-\frac{1}{2}\left\{(\hat{\boldsymbol{\alpha}} - \mathbf{m}_\theta)^\top \mathbf{Q}_\theta^{-1}(\hat{\boldsymbol{\alpha}} - \mathbf{m}_\theta) + \log\det \mathbf{Q}_\theta\right\}\right].$$

Raj Rao *et al.* (2008) used simulations to illustrate the consistency and the asymptotic normality of this estimator. However, their experiments were limited to the simplest situation where $k = 2$, and they did not provide detailed justification for their experimental findings. An important difficulty with this approach is that the functionals $\mathbf{m}_\theta$ and $\mathbf{Q}_\theta$ have no explicit form. Therefore, in the case for which the PSD H is a finite mixture (1), a software package RMTool (Raj Rao 2006) was designed to compute the asymptotic covariance matrix $\mathbf{Q}_\theta$.

In this paper, we first propose a modification of the method proposed in Raj Rao *et al.* (2008). Let F be the generalized Marčenko–Pastur distribution (Marčenko & Pastur 1967; Silverstein 1995) to which converges the empirical spectral distribution of $\mathbf{S}_n$; see Section 2 for more details. The basic idea behind our method consists of finding the map

$$\theta \mapsto \alpha_j(\theta) = \int x^j dF(x) \tag{2}$$

that links the parameter θ of the PSD H to the moments of the limiting Marčenko–Pastur distribution. Because the sample moments $(\hat{\alpha}_j)$ are consistent estimators of $(\alpha_j(\theta))$, it is then natural to use the moment method for the inference of the parameters θ.

We will consider the PSD estimation problem in two different, but related, situations, as follows.

(i) We assume a full parametric framework in which the value of k is known. We propose a moment estimator $\hat{\theta}$ for the parameters $\theta = (t_j, a_j : 1 \le j \le k)$. This estimator is proved to be strongly consistent and asymptotically Gaussian.
(ii) We adopt a nonparametric framework in which the model order k is unknown. We then combine the first estimation method with a cross-validation procedure to estimate k. Simulation experiments demonstrate the consistency of the proposed procedure.

Compared with El Karoui (2008) and Raj Rao *et al.* (2008), the main contributions of this paper are the following. The proposed moment estimator is simpler. The convergence rate of this estimator (asymptotic normality) provided in the paper is novel. The cross-validation procedure for model order selection is also a new contribution. Moreover, our method can be extended to the case for which the PSD H has not a discrete but a continuous distribution, as indicated in Section 5.

The rest of the paper is organized as follows. In Section 2, we recall several results from RMT that will be needed. In Section 3, we present our moment estimator assuming a known model order k, and prove its strong consistency and asymptotic normality. In Section 4, where k is unknown, we introduce a model selection procedure using moment estimators and a cross-validation scheme. Intensive experiments illustrate the consistency of the proposed procedure. Finally, in Section 5 we indicate an extension of our method to the case where the PSD H has a probability density function (PDF) with respect to the Lebesgue measure.

2. High-dimensional sample covariance matrices and Marčenko–Pastur distributions

We first review some fundamental results on the convergence of the spectral distributions of high-dimensional sample covariance matrices to the family of Marčenko–Pastur distributions. These results constitute the basis for our estimation methods.

Let c be an arbitrary positive constant and H an arbitrary probability measure on $\mathbb{R}^+$. Define on the set $\mathbb{C}^+ = \{z \in \mathbb{C} : \Im(z) > 0\}$ the map

$$g(s) = g_{c,H}(s) = -\frac{1}{s} + c\int \frac{t}{1+ts} dH(t), \quad s \in \mathbb{C}^+. \tag{3}$$

It is well known (Bai & Silverstein 2006, chapter 6) that g is a one-to-one map from $\mathbb{C}^+$ onto itself, and the inverse map $m = g^{-1}$ corresponds to the Stieltjes transform of a probability measure $F_{c,H}$ on $[0, \infty)$,

$$m(z) = \int_{\mathbb{R}} \frac{1}{x-z} dF_{c,H}(x), \quad z \in \mathbb{C}^+. \tag{4}$$

Moreover, by the inversion theorem of Stieltjes transforms, $F_{c,H}$ is unique. We call $F_{c,H}$ the generalized Marčenko–Pastur (M.P.) distribution with indexes (c, H).

Following the pioneering work of Marčenko & Pastur (1967), this family of distributions arises as the limits of the ESD of high-dimensional sample covariance matrices. A precise statement of this theorem involves the following assumptions, where $\mathbf{A}^{1/2}$ denotes any Hermitian square root of a non-negative definite Hermitian matrix $\mathbf{A}$.

Assumption (a). The sample size n and dimension of a random vector p both tend to infinity, and in such a way that $p/n \to c > 0$.

Assumption (b). There is a doubly infinite array of i.i.d. complex-valued random variables (w_{ij}), $i, j \geq 1$ satisfying

$$\mathrm{E}(w_{11}) = 0, \mathrm{E}(|w_{11}|^2) = 1, \mathrm{E}(|w_{11}|^4) < \infty,$$

such that for each p, n, letting $\mathbf{W}_n = (w_{ij})_{1\leq i\leq p, 1\leq j\leq n}$, the observation vectors can be represented as $\mathbf{x}_j = \Sigma_p^{1/2}\mathbf{w}_{\cdot j}$, where $\mathbf{w}_{\cdot j} = (w_{ij})_{1\leq i\leq p}$ denotes the jth column of $\mathbf{W}_n$.

Assumption (c). The SD H_p of Σ_p weakly converges to a probability distribution function H as $n \to \infty$.

Therefore, under Assumption (b), the sample covariance matrix takes the form $\mathbf{S}_n = \frac{1}{n}\Sigma_p^{1/2}\mathbf{W}_n\mathbf{W}_n^\top\Sigma_p^{1/2}$. Let the $n \times n$ matrix $\bar{\mathbf{S}}_n = \frac{1}{n}\mathbf{W}_n^\top\Sigma_p\mathbf{W}_n$ be referred to as the companion matrix of $\mathbf{S}_n$. The spectra of $\mathbf{S}_n$ and $\bar{\mathbf{S}}_n$ are identical except for $|n-p|$ zeros. Therefore, if F_n and $\bar{F}_n$ are the respective ESDs of $\mathbf{S}_n$ and $\bar{\mathbf{S}}_n$, we have

$$n\bar{F}_n - pF_n = (n-p)\delta_0. \tag{5}$$

Following the celebrated Marčenko–Pastur theorem, see Silverstein (1995), under Assumptions (a)–(c), almost surely, the ESD $\bar{F}_n$ of $\bar{\mathbf{S}}_n$ weakly converges, as $n \to \infty$, to the non-random generalized Marčenko–Pastur distribution $F_{c,H}$ with indexes (c, H) defined in (3)–(4).

When the Marčenko–Pastur theorem holds, it easily follows by (5) that the ESD F_n of $\mathbf{S}_n$ also has a weak limit F, satisfying

$$F_{c,H} - cF = (1-c)\delta_0. \tag{6}$$

From (3) and (6), we are able to deduce, in Section 3, the fundamental moment maps (2) for our estimation methods.

3. Estimation of the population spectral distribution *H*

We consider the problem of the estimation of a PSD H with a finite support as defined in (1). In this section, the size k of the support of H is assumed known. We aim to estimate the parameters $\theta = \{a_1, \ldots, a_k, t_1, \ldots, t_k\}$. The parameter space is defined as

$$\Theta = \left\{ \theta = (a_1, \ldots, a_k, t_1, \ldots, t_k) : a_j \geq 0, \text{the } a_j\text{s are distinct}; t_j > 0, \sum_{j=1}^{k} t_j = 1 \right\}.$$

The following lemma gives the relationship between the moments of the limiting MP distribution $F_{c,H}$ and those of H.

Lemma 1. *The moments $\gamma_j = \int x^j dF_{c,H}(x)$, $j \geq 1$ of the limiting MP distribution $F_{c,H}$ are linked to the moments $\beta_j = \int t^j dH(t)$ of the PSD H by*

$$\gamma_j = \sum c^{i_1+i_2+\cdots+i_j} (\beta_1)^{i_1} (\beta_2)^{i_2} \cdots (\beta_j)^{i_j} \phi^{(j)}_{i_1,i_2,\cdots,i_j}, \tag{7}$$

where the sum runs over the following partitions of j:

$$(i_1, \ldots, i_j) : j = i_1 + 2i_2 + \cdots + ji_j, \quad i_\ell \in \mathbb{N},$$

and $\phi^{(j)}_{i_1,i_2,\cdots,i_j}$ is the multinomial coefficient

$$\phi^{(j)}_{i_1,i_2,\cdots,i_j} = \frac{j!}{i_1! i_2! \cdots i_j! (j+1-(i_1+i_2+\cdots+i_j))!}. \tag{8}$$

This lemma is known, and can be proved using the fundamental equation (3); see also Nica & Speicher (2006, p. 143).

Furthermore, by (6), it is easily seen that the moment sequence (α_j) of F is proportional to that of $F_{c,H}$:

$$\alpha_j = \frac{1}{c}\gamma_j, \quad j \geq 1. \tag{9}$$

In conjunction with Lemma 1 and considering the first $2k-1$ moments, we have

$$(\alpha_1, \alpha_2, \ldots, \alpha_{2k-1}) = \Psi(\beta_1, \beta_2, \ldots, \beta_{2k-1}),$$

for an explicitly well-defined function Ψ. Note that, for all $j \geq 1$,

$$\beta_j = \sum_{\ell=1}^{k} t_\ell a_\ell^j.$$

Let us denote by Φ this functional restricted to the first $2k-1$ moments,

$$(\beta_1, \cdots, \beta_{2k-1}) = \Phi(\theta).$$

We then have, for an explicit function $g = \Psi \circ \Phi$,

$$(\alpha_1, \alpha_2, \ldots, \alpha_{2k-1}) = g(\theta), \tag{10}$$

which is the moments map (2) we are searching for.

The *moment estimator* $\hat{\theta}_n$ of θ is defined to be a solution of the moments equation

$$(\hat{\alpha}_1, \ldots, \hat{\alpha}_{2k-1}) = g(\theta), \quad \theta \in \Theta, \tag{11}$$

where $(\hat{\alpha}_j)$ are the empirical spectral moments of $\mathbf{S}_n$ defined in (1).

3.1. Strong consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$

The main result of the paper is the following:

Theorem 1. *Let θ_0 be the true value of θ and assume that Assumptions (a)–(c) hold with a PSD H of finite support as in (1) where the size k is known. Then,*

(i) *Almost surely, the moment estimator $\hat{\theta}_n$ exists for large n and converges to θ_0.*
(ii) *As $n \to \infty$,*

$$n(\hat{\theta}_n - \theta_0) \to N\big(\mathbf{A}_0\mathbf{M}_0, \mathbf{A}_0\Gamma_0\mathbf{A}_0^\top\big),$$

where $\mathbf{A}_0 = (\partial g/\partial\theta)^{-1}(\theta_0)$, and $\mathbf{M}_0$ and Γ_0 are respectively a $(2k-1)$ vector and a $(2k-1)\times(2k-1)$ matrix, defined in the proof below.

The proof of Theorem 1 relies on the following proposition.

Proposition 1. *For all $\theta \in \Theta$, we have*

$$\left|\frac{\partial\Psi}{\partial\beta}\right| \neq 0, \quad \left|\frac{\partial\Phi}{\partial\theta}\right| \neq 0, \quad \left|\frac{\partial g}{\partial\theta}\right| \neq 0.$$

Proof of Proposition 1. As $g = \Psi \circ \Phi$, it is enough to prove the first two inequalities.

(i) For $(\alpha_1, \ldots, \alpha_{2k-1}) = \Psi(\beta_1, \ldots, \beta_{2k-1})$: from (7), it is readily seen that $\alpha_j = \frac{1}{c}\gamma_j = \Psi_j(\beta_1, \ldots, \beta_j)$. Therefore, the Jacobian matrix $\partial\Psi/\partial\beta$ is lower-triangular. Moreover, its diagonal elements equal $\partial\Psi_j/\partial\beta_j = 1$. Hence, its determinant equals 1, which is positive.
(ii) For $(\beta_1, \cdots, \beta_{2k-1}) = \Phi(\theta)$: since

$$\beta_l(\theta) = \sum_{i=1}^{k} t_i a_i^l, \quad \sum_{i=1}^{k} t_i = 1,$$

we have

$$\frac{\partial \Phi}{\partial \theta} = \begin{pmatrix} a_1 - a_k & \dots & a_{k-1} - a_k & t_1 & \dots & t_k \\ a_1^2 - a_k^2 & \dots & a_{k-1}^2 - a_k^2 & 2t_1 a_1 & \dots & 2t_k a_k \\ \vdots & & & & & \\ a_1^{2k-1} - a_k^{2k-1} & \dots & a_{k-1}^{2k-1} - a_k^{2k-1} & (2k-1)t_1 a_1^{2k-2} & \dots & (2k-1)t_k a_k^{2k-2} \end{pmatrix}$$
$$= (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{2k-1})^\top.$$

Suppose that, for some vector $(b_1, b_2, \cdots b_{2k-1})$, we have

$$b_1\mathbf{u}_1 + b_2\mathbf{u}_2 + \cdots + b_{2k-1}\mathbf{u}_{2k-1} = 0.$$

Let $f(x) = b_1 x + b_2 x^2 + \cdots + b_{2k-1}x^{2k-1}$ be a polynomial function. We have then

$$f(a_1) - f(a_k) = 0, \dots, f(a_{k-1}) - f(a_k) = 0$$
$$t_1 f'(a_1) = 0, \quad t_2 f'(a_2) = 0, \dots, t_{k-1} f'(a_{k-1}) = 0, \quad \left(1 - \sum_{i=1}^{k-1} t_i\right) f'(a_k) = 0.$$

Therefore, $f(a_1) = \cdots = f(a_k)$ and $f'(a_1) = \cdots = f'(a_k) = 0$. Without loss of generality we can assume $a_k > a_{k-1} > \cdots > a_1$. By Rolle's theorem, we can find $\xi_1 \in (a_1, a_2)$, $\xi_2 \in (a_2, a_3), \dots, \xi_{k-1} \in (a_{k-1}, a_k)$ such that

$$f'(\xi_1) = 0, \quad f'(\xi_2) = 0, \dots, f'(\xi_{k-1}) = 0.$$

Therefore, we obtain $2k-1$ different roots for $f'(x)$, which is a polynomial of degree $2k-2$. Hence, $f = 0$; that is, $b_1 = b_2 = \cdots = b_{2k-1} = 0$. Consequently, the Jacobian matrix $\partial\Phi/\partial\theta$ is invertible and the conclusion follows. □

Proof of Theorem 1. We first use a central limit theorem (CLT) for linear statistics provided in Bai & Silverstein (2004, theorem 1.1). Under the assumption made, it holds that

$$n \begin{pmatrix} \hat{\alpha}_1 - \alpha_1 \\ \hat{\alpha}_2 - \alpha_2 \\ \vdots \\ \hat{\alpha}_{2k-1} - \alpha_{2k-1} \end{pmatrix} \to N(\mathbf{M}_0, \Gamma_0),$$

where the limiting mean vector $\mathbf{M}_0 = (m_\ell)$ and covariance matrix $\Gamma_0 = (\gamma_{\ell j})$, $1 \le \ell, j \le 2k-1$ are given in the reference above. Let $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_{2k-1})$ and $\alpha_0 = g(\theta_0)$. By Proposition 1 and the implicit function theorem, there exists a neighborhood U of θ_0 and a neighborhood V of α_0, such that g is a differomorphism from U onto V. Next, by the strong law of large numbers, $\hat{\alpha} \to \alpha_0$ almost surely and $\Pr(\hat{\alpha} \in V)$ tends to 1. Therefore, almost surely, for large n, $\hat{\theta}_{\mathbf{n}} = g^{-1}(\hat{\alpha}) \in U$ exists, and as $n \to \infty$, $\hat{\theta}_{\mathbf{n}} \to \theta_0 = g^{-1}(\alpha_0)$. Finally, the CLT follows from the Delta method.

TABLE 1

Estimates for the population spectral distribution H of order $k = 2$ with $n = 500$, $p = 100$ and 200 replications. True values $(a_1, a_2) = (5, 1)$ with t varying in $(0.2, 0.3, 0.5, 0.7, 0.8)$

	$\hat{t}$		$\hat{a}_1$		$\hat{a}_2$	
t	Mean	SD	Mean	SD	Mean	SD
0.2	0.1987	0.0032	5.0186	0.0694	1.0021	0.0040
0.3	0.2974	0.0043	5.0198	0.0504	1.0058	0.0105
0.5	0.5023	0.0076	5.0085	0.0435	0.9493	0.0540
0.7	0.6896	0.0146	5.0302	0.0562	1.0646	0.1071
0.8	0.7795	0.0359	5.0509	0.0872	1.1722	0.3397

TABLE 2

Estimates for the population spectral distribution H of order $k = 3$ with $n = 500$, $p = 100$ and 200 replications. True values $(a_1, a_2, a_3) = (10, 5, 1)$ with $(t_1, t_2, t_3) = (0.2, 0.4, 0.4)$

	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{t}_1$	$\hat{t}_2$
Mean	10.0922	4.9690	0.9526	0.1971	0.4127
SD	0.2421	0.3372	0.1332	0.0224	0.0323

3.2. Some simulation results

We first consider population spectral distributions H of order $k = 2 : H = t\delta_{a_1} + (1 - t)\delta_{a_2}$. Here, $2k - 1 = 3$ and the function g of (10) is given by

$$\alpha_1 = \beta_1, \quad \alpha_2 = \beta_2 + c\beta_1^2, \quad \alpha_3 = \beta_3 + 3c\beta_1\beta_2 + c^2\beta_1^3,$$

so that

$$\begin{aligned}
\alpha_1 &= ta_1 + (1 - t)a_2, \\
\alpha_2 &= ta_1^2 + (1 - t)a_2^2 + c\{ta_1 + (1 - t)a_2\}^2, \\
\alpha_3 &= ta_1^3 + (1 - t)a_2^3 + 3c\{ta_1 + (1 - t)a_2\}\left\{ta_1^2 + (1 - t)a_2^2\right\} + c^2\{ta_1 + (1 - t)a_2\}^3.
\end{aligned}$$

For the simulations in this section, we set $(a_1, a_2) = (5, 1)$, $n = 500$, $p = 100$, $c = 0.2$, and t varying in $(0.2, 0.3, 0.5, 0.7, 0.8)$. We use an array $\{w_{ij}\}$ of i.i.d. real $N(0, 1)$ variables and the following population covariance matrix:

$$\Sigma_p = \begin{pmatrix} 5I_{tp} & 0 \\ 0 & I_{(1-t)p} \end{pmatrix}.$$

Given a simulated sample covariance matrix, we used MATLAB to compute the sample eigenvalues and the moment estimator $\hat{\theta}_n = (\hat{t}, \hat{a}_1, \hat{a}_2)$. The statistics of $\hat{\theta}_n$ from 200 independent replications are summarized in Table 1.

Next, we consider the case $\hat{\theta}_n = (\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{t_1}, \hat{t_2}, \hat{t_3}) = (10, 5, 1, 0.2, 0.4, 0.4)$. The statistics of $\hat{\theta}_n$ from 200 independent replications are summarized in Table 2.

4. The model order estimation problem

When the model order k, that is, the size of the support of the PSD H, is unknown, we need to estimate it. Let (J_n) be an increasing sequence of positive integers and $D_n = \{\lambda_1, \ldots, \lambda_p\}$

the sample eigenvalues. Then, for any $1 \le k \le J_n$, there is a moment estimate $\hat{\theta}_n^{(k)}$ of θ as described in Section 3. A penalization-based model selection criterion will consist of minimizing in k a penalized pseudo-likelihood or objective function $U_n(\theta) = U_n(\theta, D_n)$. Let $(d_n)_{n\ge 0}$ be some sequence of positive numbers (penalization rate). We can estimate the true model order k_0 by

$$\hat{k}_n = \arg \min_{1\le k\le J_n} U_n\left(\hat{\theta}_n^{(k)}\right) + \frac{d_n}{n}k.$$

However, it is difficult to find a 'correct' penalization rate d_n. In particular, for the situation at hand, the 'observations', namely the sample eigenvalues (λ_i), are identically distributed but dependent. Therefore, we propose a procedure based on cross-validation that avoids such rate estimation difficulty.

4.1. A cross-validation selection procedure using likelihoods

Let $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_m, \mathbf{x}_{m+1}, \cdots, \mathbf{x}_n\}$ be a $p \times n$ data matrix as before, with i.i.d. random vectors $(\mathbf{x}_j)$. We first split it into a learning set $\mathbf{X}_{p,m} = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$ and a validation set $\mathbf{X}_{p,n-m} = \{\mathbf{x}_{m+1}, \cdots, \mathbf{x}_n\}$. Let $\mathbf{S}_{n,q} = \frac{1}{q}\mathbf{X}_{p,q}\mathbf{X}_{p,q}^\top$, $\bar{\mathbf{S}}_{n,q} = \frac{1}{q}\mathbf{X}_{p,q}^\top\mathbf{X}_{p,q}$, $q \in \{m, n-m\}$, and let $D_{n_m} = \{\lambda_1, \cdots, \lambda_p\}$, $D_{n_{n-m}} = \{\lambda_1', \cdots, \lambda_p'\}$ be the sample eigenvalues of $\mathbf{S}_{n_m}$ and $\mathbf{S}_{n_{n-m}}$, respectively.

Note that, for each PSD $H(\theta)$ with parameter θ and $c > 0$, we have a corresponding MP distribution $F_{c,H(\theta)}$ and a limit spectral distribution (LSD) F_θ for the sample covariance matrix satisfying the relation

$$F_\theta = \frac{1}{c}F_{c,H(\theta)} + \frac{c-1}{c}\delta_0; \tag{12}$$

see (6). Moreover, it is well known that, under Assumptions (a)–(c), the corresponding LSD F_θ has a bounded support with a density function f_θ on this support except an eventual mass at the origin (when $c > 1$).

For each model order $1 \le k \le J_n$, let $\hat{\theta}_n^{(k)}$ be the moment estimator based on D_{n_m}. Let

$$H(\hat{\theta}_n^{(k)}) \quad \text{and} \quad f_{\hat{\theta}_n^{(k)}}$$

be respectively the associated PSD estimate and the density function of the associated LSD for sample covariances. Using likelihoods on the validation set, we define the cross-validation estimate of the true model order k_0 as

$$\hat{k} = \arg \max_{1\le k\le J_n} \sum_{i=1}^{p} \log f_{\hat{\theta}_n^{(k)}}(\lambda_i'), \quad \lambda_i' \in D_{n_{n-m}}.$$

It is worth mentioning that, although the sample eigenvalues $\{\lambda_i'\}$ are not independent, as they share the same asymptotic probability density function of the form f_θ, the above likelihood function remains a valid estimating function. An additional difficulty occurs here because these density functions f_θ have no explicit expressions even when $H(\theta)$ is known. To solve this problem, we introduce below an approximation $\hat{f}_\theta(\lambda)$ for given θ and λ. Finally, using such approximated likelihoods, we define the cross-validation estimate of the model

order k_0 as

$$\hat{k} = \arg \max_{1 \le k \le J_n} = \sum_{i=1}^{p} \log \hat{f}_{\hat{\theta}_n^{(k)}}(\lambda_i'), \quad \lambda_i' \in D_{n_{n-m}}.$$

4.2. Approximated likelihood function $\hat{f}_\theta(\lambda)$

The approximation will be based on the following inversion theorem for Stieltjes transforms: for any cumulative distribution function G with a density function u on $\mathbb{R}$,

$$u(x) = \frac{1}{\pi} \lim_{b \to 0} \Im m_G(x + ib), \quad x \in \mathbb{R},$$

where $\Im m_G$ is the imaginary part of the Stieltjes transform m_G of G. Therefore, a natural estimator for $f_\theta(\lambda)$ is

$$\hat{f}_\theta(\lambda) = \frac{1}{\pi} \Im m_{F_\theta}(\lambda + ib), \quad \lambda \in \mathbb{R},$$

where $b > 0$ is some small number.

Let $z = \lambda + ib$. To compute $s = m_{F_\theta}(z)$, we first compute $\bar{s} = m_{F_{c,H(\theta)}}(z)$. Note that, by (3), we have

$$z = -\frac{1}{\bar{s}} + c \int \frac{t}{1 + t\bar{s}} \, dH_\theta(t).$$

Finally, $s(z)$ follows from the identity

$$s(z) = \frac{1}{c} \bar{s}(z) - \frac{c-1}{c} \frac{1}{z},$$

which is a direct consequence of (12).

4.3. Some simulation results

All simulations in this section were performed with $n = 1000$, $p = 100, m = 500$ and $b = 0.015$. The matrix entries $\{w_{ij}\}$ consisted of i.i.d. real $N(0, 1)$ variables. The number of independent runs was set to 200.

Case of a PSD H of order 2: We considered a true PSD of order $k_0 = 2$, $H = t\delta_{a_1} + (1 - t)\delta_{a_2}$, with $\theta_0 = (t, a_1, a_2) = (0.4, 5, 1)$. According to the support determination method of Silverstein & Choi (1995), the support of F_{θ_0} has two intervals $[0.37, 1.66] \cup [2.74, 8.38]$; see Figure 1.

For simulation, we used the following population covariance matrix:

$$\Sigma_p = \begin{pmatrix} 5\mathbf{I}_{0.4p} & \\ & \mathbf{I}_{0.6p} \end{pmatrix}.$$

Two values for the maximum of the candidate orders were considered: $J_n = 4$ and $J_n = 6$. Table 3 displays the resulting frequencies of the model order estimates $\hat{k}$ in these two cases.

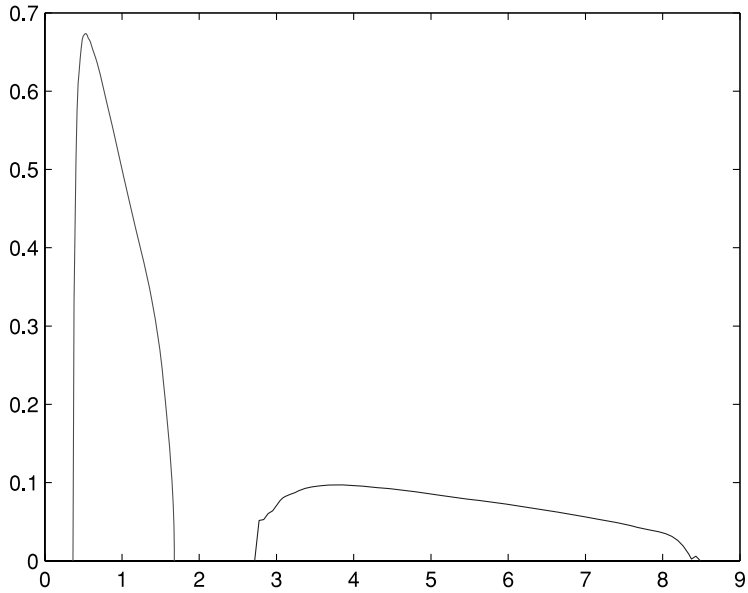

Figure 1. The probability density function f_{θ_0} of F_{θ_0} with $H = 0.4\delta_5 + 0.6\delta_1$.

TABLE 3

Distribution of the model order estimate $\hat{k}$ based on cross-validation from 200 replications. $n = 1000$, $p = 100$, $m = 500$, $b = 0.015$. True model order $k_0 = 2$. Left panel: maximum candidate order $J_n = 4$; right panel: $J_n = 6$

	$J_n = 4$				$J_n = 6$						total
$\hat{k}$	1	2	3	4	1	2	3	4	5	6	
Frequency	0	199	0	1	0	195	0	0	2	3	200

Case of a PSD H of order 3: We considered a true PSD of order $k_0 = 3$, $H = t_1\delta_{a_1} + t_2\delta_{a_2} + (1 - t_1 - t_2)\delta_{a_3}$, with $\theta_0 = (t_1, t_2, a_1, a_2, a_3) = (0.2, 0.4, 10, 5, 1)$. The support of F_{θ_0} has three intervals (see Fig. 2):

$$[0.42, 1.45] \cup [2.46, 7.55] \cup [7.59, 15.17].$$

The population covariance matrix was taken to be

$$\Sigma_p = \begin{pmatrix} 10\mathbf{I}_{0.2p} & 0 & 0 \\ 0 & 5\mathbf{I}_{0.4p} & 0 \\ 0 & 0 & \mathbf{I}_{0.4p} \end{pmatrix}.$$

Again, we considered two values, namely $J_n = 4$ and $J_n = 6$. Table 4 displays the resulting frequencies of the model order estimates $\hat{k}$ in these two cases.

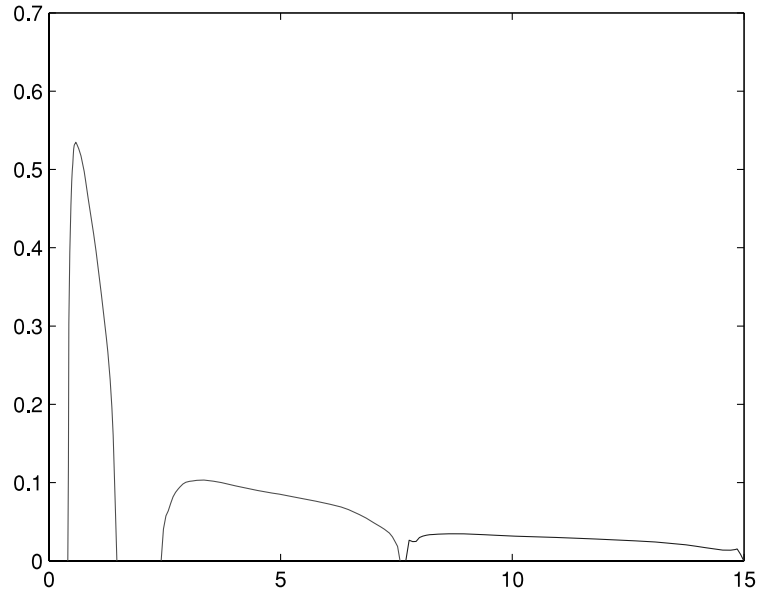

Figure 2. The probability density function f_{θ_0} of F_{θ_0} with $H = 0.2\delta_{10} + 0.4\delta_5 + 0.4\delta_1$.

TABLE 4

Distribution of the model order estimate $\hat{k}$ based on cross-validation from 200 replications. $n = 1000$, $p = 100$, $m = 500$, $b = 0.015$. True order $k_0 = 3$. Left panel: maximum candidate order $J_n = 4$; right panel: $J_n = 6$

	$J_n = 4$				$J_n = 6$						total
$\hat{k}$	1	2	3	4	1	2	3	4	5	6	
Frequency	0	1	176	23	0	0	171	15	9	5	200

5. Extension to the case where H is absolutely continuous

In this section, we indicate an extension of our method to the case where the PSD H has a probability density (with respect to Lebesgue measure):

$$dH(x) = f(x)\,dx, \quad x \in (0, \infty).$$

We assume that the unknown PDF is a continuous function, so that it has an expansion through the family of Laguerre polynomials $\{\psi_i(x)\}_{i\geq 0}$. From Szegö (1959, chapters 2, 4), this family is orthogonal with respect to the measure $e^{-x}dx$; that is, $\int \psi_i(x)\psi_j(x)e^{-x}dx = \delta_{ij}$, where δ_{ij} is the Kronecker delta. Moreover, for $m \geq 0$, $\psi_m(x) = \sum_{j=0}^{m} d_{m,j}x^j$ is a polynomial of degree n. For instance, $\psi_0(x) = 1$, $\psi_1(x) = -1 + x$.

Next, assume that f has the following finite expansion:

$$f(x) = \sum_{i=0}^{k} c_i \psi_i(x) e^{-x} = \sum_{i=0}^{k} \zeta_i x^i e^{-x},$$

TABLE 5

Estimates for a continuous population spectral distribution density f with coefficients $(\zeta_0, \zeta_1) = (0, 1)$ from 200 independent replications with $n = 500$, $p = 100$. The estimates of the distance $d = \int(\hat{f} - f)^2 dx$ are also given

	ζ_0	ζ_1	d
Mean	0.0606	0.9394	9.7736×10^{-4}
SD	0.0154	0.0154	4.5332×10^{-4}

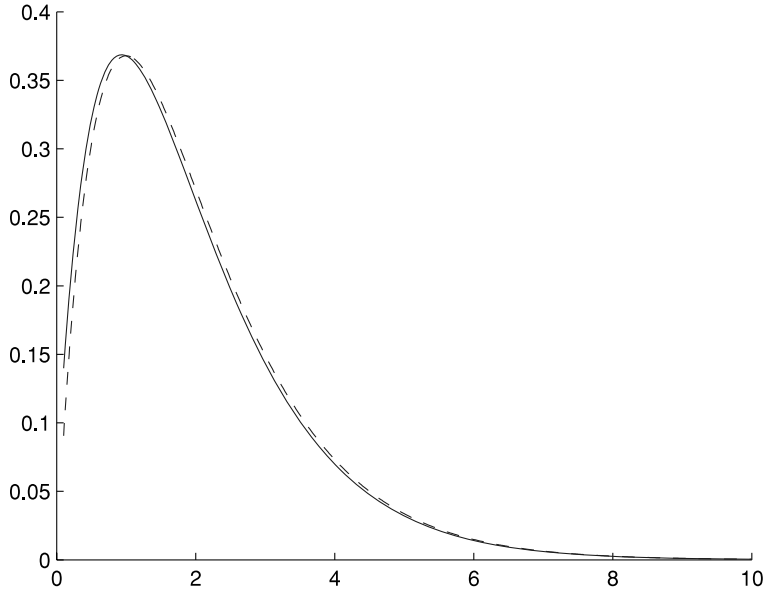

Figure 3. The solid line represents the true PSD density $f(x) = xe^{-x}$ and the dashed line represents an averaged spectral density estimate $\bar{f}(x) = (0.0606 + 0.9394x)e^{-x}$ from 200 independent replications.

where k is assumed known. The family of coefficients $\{c_i\}$ are the solution to the system

$$c_i = \int \psi_i(x) f(x) dx = \sum_{j=0}^{i} d_{ij} \int x^j f(x) dx = \sum_{j=0}^{i} d_{ij} \beta_j, \quad i = 0, 1, \ldots,$$

where β_j is the jth moment of the PSD H.

Recall the sample spectral moments $\{\hat{\alpha}_j\}$ in (1). We first obtain the estimators $\{\hat{\beta}_j\}$ of $\{\beta_j\}$ through (7)–(9). An estimator of f readily follows by taking

$$\hat{f}(x) = \sum_{i=0}^{k} \hat{c}_i \psi_i(x) e^{-x} = \sum_{i=0}^{k} \hat{\zeta}_i x^i e^{-x},$$

where $\hat{c}_i = \sum_{j=0}^{i} d_{ij} \hat{\beta}_j$.

To illustrate this extension, let us first consider a gamma distribution with shape parameter 2 and scale parameter 1; that is, $f(x) = xe^{-x}$. Therefore, $k = 1$, $(\zeta_0, \zeta_1) = (0, 1)$. Statistics for the estimates $(\hat{\zeta}_0, \hat{\zeta}_1)$ and the L^2 distance $d = \int(\hat{f} - f)^2 dx$ from 200

TABLE 6

Estimates for a continuous population spectral distribution density f with coefficients $(\zeta_0, \zeta_1, \zeta_2, \zeta_3) = (0, 1/9, 1/9, 1/9)$ from 200 independent replications with $n = 500$, $p = 100$. The estimates of the distance $d = \int(\hat{f} - f)^2 dx$ are also given

	ζ_0	ζ_1	ζ_2	ζ_3	d
Mean	0.0842	−0.2762	0.4529	0.0477	0.0049
SD	0.1430	0.4916	0.2836	0.0366	0.0048

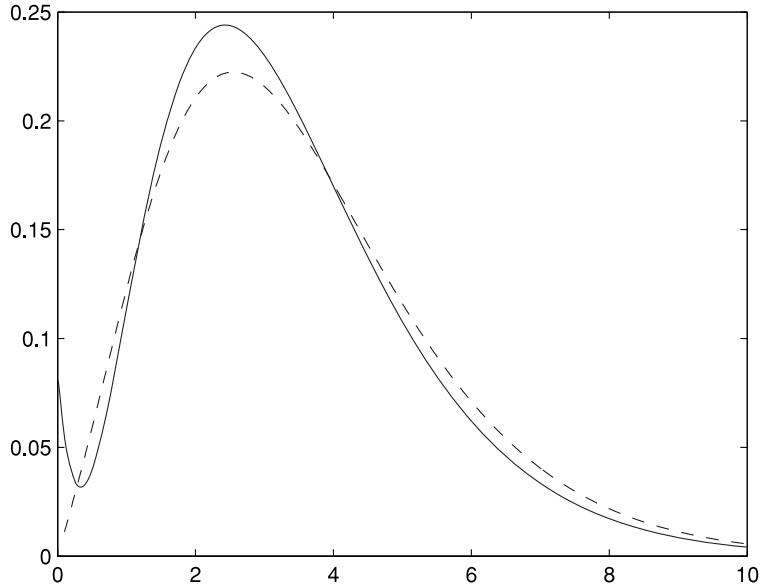

Figure 4. The solid line represents the true population spectral density $f(x) = 1/9(x + x^2 + x^3)e^{-x}$ and the dashed line represents an averaged spectral density estimate $\bar{f}(x) = (0.0842 - 0.2762x + 0.4529x^2 + 0.0477x^3)e^{-x}$ from 200 independent replications.

independent replications are summarized in Table 5. The averaged spectral density estimate $\bar{f}(x) = (0.0606 + 0.9394x)e^{-x}$, namely with the averages of the parameter estimates, is displayed in Figure 3.

Next, we consider another PDF, $f(x) = 1/9(x + x^2 + x^3)e^{-x}$. Here, $k = 3$ and $(\zeta_0, \zeta_1, \zeta_2, \zeta_3) = (0, 1/9, 1/9, 1/9)$. Analogous statistics are given in Table 6 and the averaged spectral density estimate is displayed in Figure 4. This density looks more difficult to estimate, as the parameter estimates have not really converged yet with the dimensions used. However, the L^2 distances d remain very small, which indicates the consistency of the estimators, although the convergence is quite slow in this case.

6. Concluding remarks

For the case when the parametric dimension k of the population spectral distribution H is known, we have proved the strong consistency of the proposed moment estimator as well as its asymptotic normality. These convergence results are well confirmed by the simulation

experiments of Section 3 and Section 5 for both discrete and continuous cases. This moment estimator also has an attractive simplicity.

When the model order is unknown, the proposed cross-validation procedure using approximated likelihoods from Stieltjes inversion performs correctly, as demonstrated by simulations in Section 4. A proof of these empirical findings remains to be found. This problem seems challenging, as the sample eigenvalues are dependent observations.

References

BAI, Z.D. & SILVERSTEIN, J.W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32**, 553–605.

BAI, Z.D. & SILVERSTEIN, J.W. (2006). *Spectral Analysis of Large Dimensional Random Matrices*. Beijing: Science Press.

DEMPSTER, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist.* **29**, 995–1010.

DEMPSTER, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics* **16**, 41–50.

EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36**, 2757–2790.

MARČENKO, V. & PASTUR, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb* **1**, 457–483.

NICA, A. & SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability*. New York: Cambridge University Press.

RAJ RAO, N. (2006). *RMTool - A Random Matrix Calculator in MATLAB*. Available from URL: http://www.eecs.umich.edu/∼rajnrao/rmtool/.

RAJ RAO, N., MINGO, J. A., SPEICHER, R. & EDELMAN, A. (2008). Statistical eigen-inference from large Wishart matrices. *Ann. Statist.* **36**, 2850–2885.

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

SILVERSTEIN, J. W. & CHOI, S.-I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.* **54**, 295–309.

SZEGÖ, G. (1959). *Orthogonal Polynomials*. New York: American Mathematical Society.