Journal of Multivariate Analysis 102 (2011) 1388–1398

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On a model selection problem from high-dimensional sample covariance matrices

J. Chen [a], B. Delyon [b], J.-F. Yao [b,c,*]

[a] Northeast Normal University, China
[b] Université de Rennes 1, France
[c] The University of Hong Kong, China

ARTICLE INFO

Article history:
Received 14 April 2010
Available online 18 May 2011

AMS 2000 subject classifications:
primary 62H25
62E20
secondary 60F05
15A52

Keywords:
Order selection
Cross-validation
Large sample covariance matrices
High-dimensional data
Marčenko–Pastur distribution

ABSTRACT

Modern random matrix theory indicates that when the population size p is not negligible with respect to the sample size n, the sample covariance matrices demonstrate significant deviations from the population covariance matrices. In order to recover the characteristics of the population covariance matrices from the observed sample covariance matrices, several recent solutions are proposed when the order of the underlying population spectral distribution is known. In this paper, we deal with the underlying order selection problem and propose a solution based on the cross-validation principle. We prove the consistency of the proposed procedure.

1. Introduction

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sequence of i.i.d. zero-mean random vectors in $\mathbb{R}^p$ or $\mathbb{C}^p$, with a common population covariance matrix Σ_p. When the population size p is not negligible with respect to the sample size n, modern random matrix theory indicates that the sample covariance matrix

$$S_n = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^*,$$

does not approach Σ_p. Therefore, classical statistical procedures based on an approximation of Σ_p by S_n become inconsistent in such high-dimensional data situations.

To be precise, let us recall that the *spectral distribution* (SD) G^A of an $m \times m$ Hermitian matrix (or real symmetric) A is the following measure generated by the set of its eigenvalues $\{\lambda_i^A\}$,

$$G^A = \frac{1}{m}\sum_{i=1}^{m} \delta_{\lambda_i^A},$$

* Corresponding author.
E-mail addresses: jiaqi.chen@univ-rennes1.fr (J. Chen), bernard.delyon@univ-rennes1.fr (B. Delyon), jeffyao@hku.hk (J.-F. Yao).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.05.005

where δ_b denotes the Dirac point measure at b. Let $(\sigma_i)_{1\le i\le p}$ be the p eigenvalues of the population covariance matrix Σ_p. We are particularly interested in the following SD

$$H_p := G^{\Sigma_p} = \frac{1}{p}\sum_{i=1}^{p}\delta_{\sigma_i}.$$

Following the point of view of random matrix theory, both sizes p and n will grow to infinity. It is then natural to assume that H_p weakly converges to a limiting distribution H when $p \to \infty$. We refer this limiting SD H as the *population spectral distribution* (PSD) of the observation model.

The main observation is that under reasonable assumptions, when both dimensions p and n become large at a proportional rate say c, almost surely, the (random) SD G^{S_n} of the sample covariance matrix S_n will converge almost surely and weakly to a deterministic distribution F, called limiting spectral distribution (LSD). Naturally this LSD F depends on the PSD H, but in general this relationship is complex and has no explicit form. The only exception is the case where all the population eigenvalues (σ_i) are unit, i.e. $H = \delta_1$; the LSD F is then explicitly known as the Marčenko–Pastur distribution with an explicit density function. For a general PSD H, this relationship is expressed via an implicit equation; see Section 2, Eq. (2).

An important question here is the recovering of the PSD H (or H_p) from the sample covariance matrix S_n. This question has a central importance in several popular statistical methodologies like Principal Component Analysis ([5]), Kalman filtering or Independent Component Analysis which all rely on an efficient estimation of some population covariance matrices.

Recently, El Karoui [4] has proposed a variational and nonparametric approach to this problem based on an appropriate distance function using the Marčenko–Pastur equation (2) below and a large dictionary made with base density functions and Dirac point masses. The proposed estimator is proved consistent in a nonparametric estimation sense assuming both the dictionary size and the number of observations n tend to infinity. However, no result on the convergence rate of the estimator, e.g. a central limit theorem, is given.

In another important work [7], the authors propose to use a suitable set of empirical moments, say the first q moments,

$$\widehat{\alpha}_k := \frac{1}{p}\mathrm{tr}S_n^k = \frac{1}{p}\sum_{i=1}^{p}\lambda_i^k, \quad k = 1, \ldots, q, \tag{1}$$

where (λ_ℓ) are the eigenvalues of S_n (assuming $p \le n$). Here a pure parametric approach is adopted: one assumes that the PSD depends on a set of real parameters θ: $H = H(\theta)$. To give a typical example, let the PSD be a mixture of two values a_1 and a_2 with respective weights t and $1-t$ $(0 < t < 1)$. For a given dimension p the population covariance matrix Σ_p will have approximately $[pt]$ eigenvalues equal to a_1 and $[p(1-t)]$ others equal to a_2. In this situation, the PSD H depends on three parameters a_1, a_2 and t. For more details on this example, we refer the reader to Section 1.1 of [7].

Therefore, when $n \to \infty$ and under appropriate normalization, the sample moments $(\widehat{\alpha}_k)$ will have a Gaussian limiting distribution with asymptotic mean and variance $\{m_\theta,\ Q_\theta\}$ which are functions of the (unknown) parameters θ. In [7], the authors propose an estimator $\widehat{\theta}_R$ of the parameters by maximizing the Gaussian likelihood, that is letting $\widehat{\alpha} = (\widehat{\alpha}_j)_{1\le j\le q}$,

$$\widehat{\theta}_R = \arg\max_{\theta}\left[-\frac{1}{2}\left\{(\widehat{\alpha}-m_\theta)^T Q_\theta^{-1}(\widehat{\alpha}-m_\theta) + \log\det Q_\theta\right\}\right].$$

Intensive simulations illustrate the consistency and the asymptotic normality of this estimator. However, their simulation experiments are limited to simplest situations and no theoretic result are provided concerning the consistency of the estimator. An important difficulty in this approach is that the functionals m_θ and Q_θ have no explicit form.

In a recent work [2], a modification of the procedure in [7] is proposed to get a direct moment estimator based on the sample moments $(\widehat{\alpha}_j)$. Compared to [4,7], this moment estimator is simpler and robust. Moreover, the convergence rate of this estimator (asymptotic normality) is also established.

However, all the aforementioned results assume that the dimension of the parameters θ is fixed and known. The underlying problem of model selection has been discussed and illustrated by simulations in [7,2], but no formal analysis and consistency result have been proved so far. In this paper, we pursue an approach introduced in [2] based on the cross-validation (CV) principle. Note that in [2], the CV procedure is based on the likelihood function. It turns out that the lack of continuity in the likelihood function causes serious analytic difficulties for a theoretic analysis of the underlying procedure. *The main contribution of the paper* is that we have successfully modified the contrast function together with a regularization step by convolution so that the final model selection procedure can be analysed rigorously and we prove its consistency by giving meaningful nonasymptotic bounds on the achieved risk. This consistency is obtained in a wide sense where H can be an infinite mixture of Dirac masses or a continuous distribution with a continuous density function. An interesting by-product here is that when using a Cauchy kernel for regularization, the smoothed eigenvalue densities can be evaluated efficiently through Stieltjes transforms which satisfy a Marčenko–Pastur equation (Section 5).

2. A moment estimator for the population spectral distribution *H*

We first recall the moment estimator introduced in [2] which serves as a starting block for our order selection method. The following three assumptions define the precise framework of this theory. As explained in Introduction, this moment

estimator originated from [7] and was motivated as an improvement of a procedure proposed in this reference. Throughout the paper, $A^{1/2}$ stands for any Hermitian square root of a non-negative definite Hermitian matrix A.

Assumption (a). The sample and population sizes n, p both tend to infinity, and in such a way that $p/n \to c \in (0, \infty)$.

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

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

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

Assumption (c). The SD H_p of Σ_p weakly converges to a probability distribution H as $n \to \infty$. Moreover, the sequence of spectral norms $(\|\Sigma_p\|)$ is bounded.

The assumptions (a)–(c) are classical conditions for the celebrated Marčenko–Pastur theorem ([6], see also [1]). More precisely, under these assumptions, it holds that almost surely, the empirical SD G^{S_n} of S_n, weakly converges, as $n \to \infty$, to the (nonrandom) generalized Marčenko–Pastur distribution F which in particular depends on c and H. It is well known that the LSD F has a bounded support with a density function f on this support except an eventual mass at the origin (when $c > 1$).

Note also that under assumption (b), the sample covariance matrix takes the form

$$S_n = \frac{1}{n} \Sigma_p^{1/2} W_n W_n^* \Sigma_p^{1/2}.$$

This representation form and the assumed boundedness of the spectral norms $(\|\Sigma_p\|)$ in assumption (c) will be explicitly used later in the main Theorem 4.1.

Unfortunately, except the simplest case where $H \equiv \delta_1$, the above LSD F has no explicit form. In the general case, F is characterized as follows. Let $s(z)$ denote the Stieltjes transform of $F_* := cF + (1 - c)\delta_0$, which is an one-to-one map defined on the upper half-complex plane $\mathbb{C}^+ = \{z \in \mathbb{C} : \Im(z) > 0\}$. This transform satisfies the following fundamental Marčenko–Pastur equation:

$$z = -\frac{1}{s(z)} + c \int \frac{t}{1 + ts(z)} dH(t), \quad z \in \mathbb{C}^+. \tag{2}$$

In [2] (see also [7]), a moment estimator of θ is introduced as follows. Let (α_j) and (β_j) be the sequences of the moments of F and H, respectively. A fundamental consequence of Marčenko–Pastur equation (2) is that for any $N \geq 1$, there is an one-to-one and explicitly known map Ψ_N which links both sets of N first moments:

$$(\alpha_1, \alpha_2, \ldots, \alpha_N) = \Psi_N(\beta_1, \beta_2, \ldots, \beta_N). \tag{3}$$

For the precise definition of Ψ_N, we refer to Refs. [2,7]. Assume that the unknown PSD H depend on k parameters $\theta = (\theta_1, \ldots, \theta_k)$ belonging to a k-dimensional real parameter space Θ. Let $F(\theta)$ thus denote the associated LSD and f_θ its density function (all density functions are with respect to the Lebesgue measure throughout the paper). For example, in the discrete case, we are often considering a family of finite mixture of Dirac masses

$$H(\theta) = \sum_{\ell=1}^{m} t_\ell \delta_{a_\ell},$$

with $a_\ell \geq 0$, $t_\ell \geq 0$ and $\sum t_\ell = 1$. Here we have $k = 2m - 1$ parameters (a_ℓ) and (t_ℓ). Note that such a PSD H corresponds, for a given dimension p, to the situation where the population eigenvalues (σ_i) of the covariance matrix Σ_p coincide with the a_ℓ's whose multiplicity number approximately equals $[t_\ell p]$.

In general, given a parametric form $H(\theta)$, we can define an explicit map which links the k parameters to the k first moments of H:

$$(\beta_1, \ldots, \beta_k) = \Phi(\theta).$$

For instance in the previous discrete case, we have simply for any $j \geq 1$,

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

For the general case, we have for an explicit function $\Xi_k = \Psi_k \circ \Phi$

$$(\alpha_1, \alpha_2, \ldots, \alpha_k) = \Xi_k(\theta). \tag{4}$$

Recalling the empirical moments $(\widehat{\alpha}_j)$ defined in (1), the *moment estimator* $\widehat{\theta}_n$ of the parameter θ is defined to be any solution of the moment equation

$$(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_k) = \Xi_k(\theta), \quad \theta \in \Theta. \tag{5}$$

When the model order k is known and under suitable regularity conditions, the strong consistency and the asymptotic normality of the moment estimator $\widehat{\theta}_n$ are established in [2].

3. A cross-validation procedure to estimate the model order

When the model order k, i.e. the number of the parameters which determine the PSD H, is unknown, also we need to estimate it from the data. A main difficulty here is that the data, namely the sample eigenvalues (λ_j) are dependent observations. In this work, we propose an order selection procedure based on the cross-validation. From now on, we denote by H_0 the true PSD to be estimated, and by F_0 and $g := f_0$ the associated LSD and its density function, respectively.

Let (J_n) be an increasing sequence of positive integers and $\{\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1}, \ldots, \mathbf{x}_{m+n}\}$ a sample of i.i.d. random vectors as before. We first split it to a training set $\mathbf{X}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, and a validation set $\mathbf{X}_2 = \{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+m}\}$. Let

$$S_1 = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j\mathbf{x}_j^*, \qquad S_2 = \frac{1}{m}\sum_{j=n+1}^{n+m} \mathbf{x}_j\mathbf{x}_j^*,$$

be the associated sample covariance matrices, with eigenvalues $D_1 = \{\lambda_1, \ldots, \lambda_p\}$ and $D_2 = \{\lambda'_1, \ldots, \lambda'_p\}$, respectively.

To simplify the presentation, we will hereafter assume that both training and validation sets have an equal size $m = n$ although the general case with $m \neq n$ can be handled exactly in the same manner.

For each $1 \leq k \leq J_n$, let $\widehat{\theta}_n^{(k)}$ be the moment estimator based on D_1, that is from the learning set $\mathbf{X}_1$ and model dimension k, as recalled in Section 2. Let $H(\widehat{\theta}_n^{(k)})$ be the associated PSD estimate, $f_{\widehat{\theta}_n^{(k)}}$ the density function of the associated LSD estimate $F_{\widehat{\theta}_n^{(k)}}$. We need to choose an appropriate contrast function $K(f)$ on the validation set to estimate the order k_0 of the true PSD H_0. Naturally, we consider the likelihood method and we may obtain the estimation of k_0 as follows:

$$\widehat{k}_1 = \arg\max_{1\leq k\leq J_n} \sum_{i=1}^{p} \log f_{\widehat{\theta}_n^{(k)}}(\lambda'_i), \quad \lambda'_i \in D_2. \tag{6}$$

An additional difficulty happens here because the density functions f_θ have no explicit expressions even when $H(\theta)$ is known. To solve this problem, we use an approximation $\widehat{f}_\theta(\lambda')$ for any given θ and λ' introduced in [2] and based on the inversion formula of Stieltjes transform; see also Eq. (13) below.

The likelihood-based selection rule (6) is tested on several simulation experiments leading to relatively satisfying results (see [2]). However, for a theoretical analysis of this rule, we have a serious difficulty when some of the sample eigenvalues λ'_i from the validation set approach the boundary of the support of the LSD estimate $F_{\widehat{\theta}_n^{(k)}}$. Indeed, at these values, the log-likelihoods $\log f_{\widehat{\theta}_n^{(k)}}(\lambda'_i)$ become unbounded. To overcome such analytical difficulty, we are led to substitute a smoother contrast function for the likelihood function. A first idea is to use the following least-squares function

$$K_n^0(f) = \frac{1}{2}\int f(x)^2 dx - \frac{1}{p}\sum_{i=1}^{p} f(\lambda'_i), \quad \lambda'_i \in D_2. \tag{7}$$

Note that this usual L_2 distance is widely used in the literature of nonparametric density estimation.

Actually, this is a valid contrast function since its mean equals

$$K^0(f) = EK_n^0(f) = \int \left(\frac{1}{2}f(x)^2 - f(x)g(x)\right) dx,$$

and we have

$$K^0(f) - K^0(g) = \frac{1}{2}\|f - g\|_2^2.$$

We can then propose a new cross-validation rule:

$$\widehat{k}_2 = \arg\min_{1\leq k\leq J_n} K_n^0(f_{\widehat{\theta}_n^{(k)}}). \tag{8}$$

Unfortunately, a Marčenko–Pastur density function f lacks smoothness at the boundary. Indeed, near a boundary point a, $f(\lambda)$ behaves as $\sqrt{|\lambda - a|}$ ([6,8]). Therefore, f is not differentiable at boundary. This makes the analysis of the selection rule (7)–(8) difficult.

Our solution to this problem is to use a smoothed version of f in (7). Let φ be a smooth kernel function. We propose to use the following contrast function

$$K_n(f) = \frac{1}{2}\int f_\varphi(x)^2 dx - \frac{1}{p}\sum_{i=1}^{p} f_{\varphi\check{\varphi}}(\lambda'_i), \quad \lambda'_i \in D_2, \tag{9}$$

where $f_\varphi = f * \varphi$, $f_{\varphi\check{\varphi}} = f * \varphi * \check{\varphi}$, $\check{\varphi}(x) = \varphi(-x)$. This is again a valid contrast function since simple computations prove that its mean function $K(f) = EK_n(f)$ satisfies

$$K(f) - K(g) = \frac{1}{2}\|f_\varphi - g_\varphi\|_2^2.$$

Finally, here is the cross-validation rule we introduce in this paper

$$\widehat{k} = \arg\min_{1\le k\le J_n} K_n(f_{\widehat{\theta}_n^{(k)}}). \tag{10}$$

With this order estimate, we have

$$\widehat{g} = f_{\widehat{\theta}_n^{(\widehat{k})}}, \tag{11}$$

as the final estimate of the density $g = f_0$ for the true LSD F_0.

4. Consistency of the cross-validation procedure

Define the risk function

$$r(f) = \frac{1}{2}\|(f-g)*\varphi\|_2^2$$

and g is the density function of the true LSD F_0. The main result of the paper is the following.

Theorem 4.1. *Assume that assumptions* (a)–(b)–(c) *hold with the matrix entries* $\{w_{ij}\}$ *uniformly bounded by a constant* κ. *Then, for the cross-validation estimate* $\widehat{g}$ *in* (11) *and any* $\varepsilon > 0$

$$(1-\varepsilon)E[r(\widehat{g})] \le \min_{1\le k\le J_n} r(f_{\widehat{\theta}_n^{(k)}}) + \alpha_0 \frac{\log(J_n)}{\varepsilon np},$$

where the expectation is conditional to D_1 *and*

$$\alpha_0 = 64a^4\left(\|\varphi'\|_2 + \frac{p}{n}a^2\|\varphi''\|_2\right)^2,$$
$$a = \kappa \sup_{p\ge 1}\|\Sigma_p^{1/2}\|.$$

To explain the content of the above theorem, let us first consider a parametric setting. Assume then there is a finite order k_0 and a true parameter value θ_0 at this order such that the unknown PSD is $H = H(\theta_0)$. Therefore, there is a true LSD density $g = f_{\theta_0}$. According to [2] (see also Section 2), the moment estimator $\widehat{\theta}_n^{(k_0)}$ at the order k_0 has an asymptotic Gaussian distribution. In particular,

$$\widehat{\theta}_n^{(k_0)} - \theta_0 = O_P\left(\frac{1}{\sqrt{np}}\right).$$

It follows that under reasonable continuity conditions on the map $\theta \mapsto f_\theta$, we will have

$$r(f_{\widehat{\theta}_n^{(k_0)}}) = O_P\left(\frac{1}{np}\right).$$

Therefore, if this true order k_0 were known, one would use this value of k_0 and would not get, for the minimum risk $\min_k r(\widehat{f_{\widehat{\theta}_n^{(k)}}})$, better than the order $(np)^{-1}$. The additional logarithmic term in the theorem above is thus a standard adaptation cost which typically behaves as $\log(np)$ when e.g. J_n is some power of np.

Otherwise, we run into a nonparametric framework, $g \ne f_{\theta^{(k)}}$ for any finite-dimensional parameter θ_k, and the minimum risk term could not be expected to be smaller than $(np)^{-\alpha}$ for some $\alpha < 1$, and the additional logarithmic term becomes negligible.

The proof of Theorem 4.1 relies on the following concentration inequality for eigenvalues of random matrices proposed in [3]. Let $\|x\|$ be the Euclidean norm on $\mathbb{R}^d$ and $\|M\|$ the associated operator norm for a $d\times d$ matrix M.

Proposition 4.1 (*[3]*). *Let* B *be a* $p\times p$ *deterministic matrix,* $Z = (Z_{ij})$, $1\le i\le p$, $1\le j\le n$ *be a matrix of random independent entries, and set* $M = \frac{1}{n}BZZ^*B^*$. *Let* $\lambda \mapsto q(\lambda)$ *be a differentiable symmetric function on* $\mathbb{R}^p$ *and define the random variable* $W = q(\lambda) = q(\lambda_1,\dots\lambda_p)$ *where* $(\lambda_1,\dots\lambda_p)$ *is the vector of the eigenvalues of* M. *Then*

$$E\left[e^{W-E[W]}\right] \le \exp\left(\frac{64p}{n}a^4\left(\gamma_1 + \frac{p}{n}a^2\gamma_2\right)^2\right), \tag{12}$$

where

$$a = \|B\|\sup_{ij}\|Z_{ij}\|_\infty, \qquad \gamma_1 = \sup_{k,\lambda}\left|\frac{\partial q}{\partial\lambda_k}(\lambda)\right|,$$
$$\gamma_2 = \sup_\lambda\|\nabla^2 q(\lambda)\|, \quad \nabla^2 q(\lambda) := \left(\frac{\partial^2 q}{\partial\lambda_j\partial\lambda_k}(\lambda)\right)_{1\le j,k\le p}.$$

Proof of Theorem 4.1. With the empirical contrast function K_n defined in (9), we have

$$R(f) := K_n(f) - K_n(g) = \frac{1}{2}\int \left\{f_\varphi(x)^2 - g_\varphi(x)^2\right\} dx - \frac{1}{p}\sum_{i=1}^{p} \left\{f_{\varphi\check{\varphi}} - g_{\varphi\check{\varphi}}\right\}(\lambda_i'), \quad \lambda_i' \in D_2,$$

and

$$\begin{aligned} r(f) = E[R(f)] &= \int \left(\frac{1}{2}f_\varphi(x)^2 - f_\varphi(x)g_\varphi(x) - \frac{1}{2}g_\varphi(x)^2 + g_\varphi(x)^2\right) dx \\ &= \frac{1}{2}\int (f_\varphi(x) - g_\varphi(x))^2 dx. \end{aligned}$$

We are going to apply Proposition 4.1 to the random variable $W = -cR(f)$ with some positive constant $c > 0$ and the sample covariance matrix $S_2 = \frac{1}{n}\Sigma_p^{1/2} W_n W_n^* \Sigma_p^{1/2}$. As the entries (w_{ij}) of W_n are bounded by κ, we can take for the constant a

$$a = \kappa \sup_{p\geq 1} \|\Sigma_p^{1/2}\|.$$

Next, we have

$$q(\lambda') = q(\lambda_1', \ldots, \lambda_p') = -cR(f),$$

so that

$$\begin{aligned} \frac{\partial q}{\partial \lambda_k'}(\lambda') &= \frac{c}{p}(f'_{\varphi\check{\varphi}} - g'_{\varphi\check{\varphi}})(\lambda'), \\ \frac{\partial^2 q}{\partial \lambda_j' \partial \lambda_k'}(\lambda') &= \frac{c}{p}(f''_{\varphi\check{\varphi}} - g''_{\varphi\check{\varphi}})(\lambda_j')1_{\{j=k\}}. \end{aligned}$$

Hence,

$$\begin{aligned} \sup_{k,\lambda'} \left|\frac{\partial q}{\partial \lambda_k'}(\lambda')\right| &\leq \frac{c}{p}\|f'_{\varphi\check{\varphi}} - g'_{\varphi\check{\varphi}}\|_\infty =: \frac{c}{p}\gamma_1(f), \\ \sup_{\lambda'} \|\nabla^2_{\lambda'}\| &\leq \frac{c}{p}\|f''_{\varphi\check{\varphi}} - g''_{\varphi\check{\varphi}}\|_\infty =: \frac{c}{p}\gamma_2(f), \end{aligned}$$

where we have denoted the infinite norms by $\gamma_1(f)$ and $\gamma_2(f)$. Applying Proposition 4.1, we obtain for any f

$$E\left[e^{-cR(f)+cr(f)}\right] \leq \exp\left(\tfrac{64p}{n}\, a^4c^2\left(2\gamma_1(f) + \tfrac{p}{n}a^2\gamma_2(f)\right)^2\right).$$

Next we need to bound the two infinite norms by the risk function $r(f)$. Notice that for any $h \in L_2$, one has

$$\|(h * \check{\varphi})'\|_\infty = \|h * (\check{\varphi}')\|_\infty \leq \|h\|_2\|\varphi'\|_2,$$

and similarly

$$\|(h * \check{\varphi})''\|_\infty \leq \|h\|_2\|\varphi''\|_2,$$

and applying these inequalities with $h = (f - g) * \varphi$, we get

$$\begin{aligned} \gamma_1(f) &\leq \|\varphi'\|_2\|f_\varphi - g_\varphi\|_2 = \|\varphi'\|_2\sqrt{r(f)}, \\ \gamma_2(f) &\leq \|\varphi''\|_2\|f_\varphi - g_\varphi\|_2 = \|\varphi''\|_2\sqrt{r(f)}. \end{aligned}$$

Hence

$$\begin{aligned} E\left[e^{-cR(f)+cr(f)}\right] &\leq \exp\left(\frac{64}{np}\, a^4c^2\left(\|\varphi'\|_2 + \frac{p}{n}a^2\|\varphi''\|_2\right)^2 r(f)\right) \\ &= \exp\left(\frac{\alpha_0}{np}c^2 r(f)\right), \end{aligned}$$

with

$$\alpha_0 := 64a^4\left(\|\varphi'\|_2 + \frac{p}{n}a^2\|\varphi''\|_2\right)^2.$$

This inequality is true for any of the $f_{\widehat{\theta}_n^{(k)}}$, $k \le J_n$ and we remind the reader that the expectation is taken over the validation data conditionally to the training data D_1. We recall that $\widehat{k} = \widehat{k}(\omega)$ is the minimizer of $K_n(f_{\widehat{\theta}_n^{(k)}})$ which is also the minimizer of $R(f_{\widehat{\theta}_n^{(k)}})$. If we set

$$R_k = R(f_{\widehat{\theta}_n^{(k)}}),$$

$\widehat{k} = \widehat{k}(\omega)$ is the random index such that

$$R_{\widehat{k}} \le R_k, \quad k \le J_n.$$

Let m denote the index j which attains the minimum of r_j, $1 \le j \le J_n$; this is the best possible choice. For any $0 < \lambda \le 1$:

$$\begin{aligned}
\lambda E[r_{\widehat{k}}] &\le E[\lambda r_{\widehat{k}} + R_m - R_{\widehat{k}}] \\
&= r_m + E[\lambda r_{\widehat{k}} - R_{\widehat{k}}] \\
&\le r_m + c^{-1} \log E\left[e^{c(\lambda r_{\widehat{k}} - R_{\widehat{k}})}\right] \\
&\le r_m + c^{-1} \log E\left[\sum_j e^{c(\lambda r_j - R_j)}\right] \\
&\le r_m + c^{-1} \log J_n \sup_j e^{c\lambda r_j} E[e^{-cR_j}] \\
&\le r_m + c^{-1} \log J_n \sup_j e^{-c(1-\lambda)r_j} e^{\frac{\alpha_0 c^2}{np} r_j} \\
&= r_m + c^{-1} \log J_n + c \sup_j \left(-(1-\lambda)r_j + \frac{\alpha_0 c}{np} r_j\right),
\end{aligned}$$

where $1 \le j \le J_n$. By taking $\lambda = 1 - c\alpha_0/(np)$,

$$(1 - c\alpha/(np))E[r_{\widehat{k}}] \le \min_j r_j + \frac{\log(J_n)}{c}.$$

We now take $c = \varepsilon np/\alpha_0$,

$$(1-\varepsilon)E[r_{\widehat{k}}] \le \min_j r_j + \alpha_0 \frac{\log(J_n)}{\varepsilon np}.$$

The proof is complete. □

5. Implementation of the procedure with a canonical choice of φ

This section is aimed to describe the practical implementation of our procedure. First of all we need to choose a smoothing kernel φ. An amazing and important fact here is that there is a very natural choice for φ and it seems to us that any other choice will result in considerable computing difficulties for the proposed cross-validation procedure.

Indeed, the family of Cauchy densities

$$C_\eta(x) = \frac{\eta}{\pi(x^2 + \eta^2)}, \quad x \in \mathbb{R},$$

where $\eta > 0$ is a parameter, and is intimately related to the Stieltjes transformation. Given an LSD F with a density function f, let us recall its Stieltjes transform

$$s_F(z) = \int \frac{1}{\lambda - z} dF(\lambda), \quad z \in \mathbb{C}^+.$$

It is easy to see by letting $z = x + i\eta$ with $x \in \mathbb{R}$ and $\eta > 0$ that

$$\frac{1}{\pi} \Im s_F(x + i\eta) = \frac{1}{\pi} \int \frac{\eta f(\lambda)}{(x-\lambda)^2 + \eta^2} d\lambda = f * C_\eta(x).$$

Since (C_η) is a regular approximation of the unity (for the convolution operator) when $\eta \to 0$, we get immediately the following Stieltjes inversion formula: for any $x \in \mathbb{R}$,

$$f(x) = \lim_{\eta \to 0} \Im s_F(x + i\eta). \tag{13}$$

Coming back to the smoothed contrast function $K_n(f)$ in (9), there is then a canonical choice $\varphi = C_\eta$ for some given width $\eta > 0$, since the values of $s_F(x + i\eta)$ can be obtained through the Marčenko–Pastur equation (2) for any given PSD H and the associated LSD F.

Let us summarize all the steps of our cross-validation method as follows.

1. First split the data into the training and validation sets as described before.
2. Compute then the eigenvalues $D_1 = \{\lambda_i\}$ and $D_2 = \{\lambda_j'\}$ from the associated sample covariance matrices.
3. Choose a small positive value η for the Cauchy kernel $\varphi = C_\eta$.
4. Choose J_n as an a priori upper bound for the unknown model order.

Next for each $0 \le k \le J_n$, we obtain the moment estimator $f_{\widehat{\theta}_n^{(k)}}$ based on D_1. We then compute its CV contrast value based on D_2 using the kernel C_η:

$$K_n(f_{\widehat{\theta}_n^{(k)}}) = \frac{1}{2}\int (f_{\widehat{\theta}_n^{(k)}} * C_\eta)^2(x)dx - \frac{1}{p}\sum_{j=1}^{p} f_{\widehat{\theta}_n^{(k)}} * C_\eta * \check{C}_\eta(\lambda_i'),$$

by observing the following property:

$$f_{\widehat{\theta}_n^{(k)}} * C_\eta(x) = \frac{1}{\pi}\Im\widehat{s}_{F_{\theta_n^{(k)}}}(x + i\eta).$$

Here, the estimator $\widehat{s}_{F_{\theta_n^{(k)}}}$ of s is calculated using the equation

$$s = \int \frac{1}{t(1 - c - czs) - z} dH_{\theta_n^{(k)}}(t),$$

which is another well-known relation on the Stieltjes transforms equivalent to Eq. (2) (see [1] for more details).

Furthermore as $C_\eta * \check{C}_\eta = C_\eta * C_\eta = C_{2\eta}$, we have $f_{\widehat{\theta}_n^{(k)}} * C_\eta * \check{C}_\eta = f_{\widehat{\theta}_n^{(k)}} * C_{2\eta}$. Therefore substituting 2η for η in the previous computation, we are also able to evaluate the second term of the contrast function K_n.

Finally, the order estimate $\widehat{k}$ is picked up as the one minimizing these K_n values.

6. Extension to the case where H is absolutely continuous

In this section, we indicate an extension of our estimation method as well as the cross-validation procedure for order selection to the case where the PSD H has a density (with respect to Lebesgue measure):

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

We assume that the unknown density function h is a continuous function, so that it has an expansion through the family of Laguerre polynomials $\{\psi_i(x)\}_{i\ge 0}$ [9, Chap.2,4]:

$$h(x) = \sum_{i=1}^{\infty} c_i \psi_{i-1}(x)e^{-x} = \sum_{i=1}^{\infty} \zeta_i x^{i-1}e^{-x}.$$

The family of coefficients $\{c_i\}$ are solution to the system

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

where β_j is the jth moment of H and $\{d_{ij}\}$ a family of explicitly known constants.

Furthermore, for any given truncation order k, we can, as for the discrete case, obtain estimates $\{\widehat{\beta}_j\}_{1\le j\le k}$ of the first k moments of H through Eqs. (1) and (3). A moment estimator for the unknown PSD density h thus follows

$$\widehat{h}_k(x) = \sum_{i=1}^{k} \widehat{c}_i \psi_{i-1}(x)e^{-x}, \tag{14}$$

with

$$\widehat{c}_i = \sum_{j=1}^{i} d_{ij}\widehat{\beta}_j, \quad 1 \le i \le k.$$

Next, for selection of the truncation order k, we adapt the previous cross-validation rule (9)–(10) to the present case. We split a data set to a training set and a validation set exactly as before. Using the training set, we get, for any $1 \le k \le J_n$, a density estimate $\widehat{h}_k$ by the moment method, Eq. (14). Therefore, the order estimate is defined as

$$\widehat{k}_c = \arg\min_{1\le k\le J_n} K_n(\widehat{h}_k), \tag{15}$$

where the contrast function K_n is the one defined in (9) using the validation data.

7. Simulation results

All the simulations reported in this section use i.i.d. Gaussian variates $\{w_{ij}\}$ and the following parameters: $n = m = 500$ and $p = 100$; $\eta = 0.025$ for the discrete case and $\eta = 0.015$ for the continuous case. In the following, I_s denotes the s-dimensional identity matrix.

Table 1
Distribution of the model order estimate $\widehat{k}$ and averages of intra-class Wasserstein distances from 200 replications. $n = m = 500$, $p = 100$, $\eta = 0.025$ and $J_n = 6$. True model order $k_0 = 2$.

$\widehat{k}$	1	2	3	4	5	6	Total
Frequency	0	187	5	0	4	4	200
δ	-	0.0597	0.1297	-	0.4115	0.3365	

Table 2
Distribution of the model order estimate $\widehat{k}$ and averages of intra-class Wasserstein distances from 200 replications. $n = m = 500$, $p = 100$, $\eta = 0.025$ and $J_n = 6$. True model order $k_0 = 3$.

$\widehat{k}$	1	2	3	4	5	6	Total
Frequency	0	0	166	14	15	5	200
δ	-	-	0.3268	0.3935	0.8084	0.6860	

Table 3
Distribution of the model order estimate $\widehat{k}$ for a continuous PSD density and averages of intra-class L^1 distances from 200 replications. $n = m = 500$, $p = 100$, $\eta = 0.015$ and $J_n = 5$. True model order $k_0 = 3$.

$\widehat{k}$	1	2	3	4	5	Total
Frequency	0	0	155	7	38	200

Case of a discrete PSD H of order 2. We consider a true PSD of order $k_0 = 2$: $H_0 = t\delta_{a_1} + (1-t)\delta_{a_2}$, with $t = 0.4$ and $(a_1, a_2) = (5, 1)$. The population covariance matrix is set to be

$$\mathbf{\Sigma}_p = \begin{pmatrix} 5I_{0.4p} & 0 \\ 0 & I_{0.6p} \end{pmatrix}.$$

For order selection, we use $J_n = 6$ and repeat 200 independent experiments. The frequencies of the cross-validation model order estimates $\widehat{k}$ over the 200 replications are summarized in Table 1. Note that the last line in the table displays for each class the average δ of first-order Wasserstein distance $\mathcal{W}_1(H_0, H(\widehat{\theta}_n^{(k)}))$ (here for discrete distributions).

Case of a discrete PSD H of order 3. Next we consider a true PSD of order $k_0 = 3$: $H_0 = t_1\delta_{a_1} + t_2\delta_{a_2} + (1 - t_1 - t_2)\delta_{a_3}$, with $(t_1, t_2, a_1, a_2, a_3) = (0.2, 0.4, 10, 5, 1)$. The population covariance matrix is set to be

$$\mathbf{\Sigma}_p = \begin{pmatrix} 10I_{0.2p} & 0 & 0 \\ 0 & 5I_{0.4p} & 0 \\ 0 & 0 & I_{0.4p} \end{pmatrix}.$$

Table 2 summarizes the frequency distribution of the cross-validation order estimate $\widehat{k}$ from 200 independent replications using $J_n = 6$, and the averaged Wasserstein distance δ.

Case of a continuous PSD H. Here for the true PSD H_0, we consider a Gamma distribution with shape parameter 3 and scale parameter 1, i.e $h(x) = \frac{1}{2}x^2e^{-x}$.

Based on the cross-validation rule (15), Table 3 summarizes the frequency distribution of the cross-validation order estimate $\widehat{k}$ from 200 independent replications using $J_n = 5$, and the average of L^1 distance $\int |h(x) - \widehat{h}_{\widehat{k}}(x)|dx$ within the classes.

On the influence of the smoothing parameter η. It is not trivial to define a priori choice of the smoothing parameter η. Here we provide some empirical findings by running the previous simulation experiments over a range of values for η.

Tables 4 and 5 display the observed distributions of the order estimate $\widehat{k}$ for the two discrete cases considered above. Overall, the cross-validation procedure seems very robust against the choice of η, except for very low values like 0.0004 and 0.0005 where the criterion becomes to loss efficiency.

Effect of the population to sample ratio p/n.

Here we want to see experimentally the effect of the population to sample ratio p/n on our procedure. Table 6 reports an experiment with fixed $m = n = 500$ while increasing p from 100 to 500 and for the discrete PSD of order 2 above.

One can observe that the method becomes less accurate as p increases. A possible explanation of this is that when the ratio p/n increases to 1, the proportion of small sample eigenvalues increases near the left edge of the support. As the density function is highly increasing (its derivative equals infinity at the edges) in this area, it is expected that the density estimates used in our procedure are less accurate.

This phenomenon is also confirmed by the risk bounds given in Theorem 4.1 involving the constant α_0 which is increasing with the ratio p/n so that the estimation problem becomes harder.

Table 4
Distribution of the model order estimate $\widehat{k}$ based on cross-validation from 200 replications. $n = m = 500$, $p = 100$ with η varying in $(0.05, 0.025, 0.0125, 0.0063, 0.001, 0.0004)$ and $Q_n = \{1, 2, 3, 4, 5, 6\}$. True model order $k_0 = 2$.

$\eta \setminus \widehat{k}$	1	2	3	4	5	6	Total
0.05	0	168	7	0	14	11	200
0.025	0	187	5	0	4	4	200
0.0125	0	196	4	0	0	0	200
0.0063	0	198	1	0	0	1	200
0.001	0	182	10	3	3	2	200
0.0004	0	113	25	25	21	16	200

Table 5
Distribution of the model order estimate $\widehat{k}$ based on cross-validation from 200 replications. $n = m = 500$, $p = 100$ with η varying in $(0.05, 0.025, 0.0125, 0.0063, 0.0005)$ and $Q_n = \{1, 2, 3, 4, 5, 6\}$. True model order $k_0 = 3$.

$\eta \setminus \widehat{k}$	1	2	3	4	5	6	Total
0.05	0	0	152	15	26	11	200
0.025	0	0	166	14	15	5	200
0.0125	0	1	165	9	22	3	200
0.0063	0	1	163	10	16	10	200
0.0005	0	7	121	20	34	18	200

Table 6
Distribution of the model order estimate $\widehat{k}$ from 200 replications. $n = m = 500$, $\eta = 0.025$, $J_n = \{1, 2, 3, 4, 5, 6\}$ and p varying in $\{100, 200, 300, 400, 450, 500\}$. True model order $k_0 = 2$.

$p \setminus \widehat{k}$	1	2	3	4	5	6	total
100	0	187	5	0	4	4	200
200	0	194	0	4	2	0	200
300	0	189	7	1	2	1	200
400	0	159	19	1	19	2	200
450	0	169	9	2	16	4	200
500	3	130	16	7	37	7	200

8. Discussions

Undoubtedly in statistical problems involving high-dimensional data, we need to develop new tools to answer the question of model selection. We have proposed in this paper an order selection method using cross-validation in the specific context of determining the population spectral distribution from the observed sample covariance matrices.

In the view of the authors, several related issues merit further investigation. First, estimations based on high moments tend to be fairly unstable and there is a need for modification of the proposed parameter estimators in order to reduce this unstability. Second, our cross-validation criterion is based on a kernel smoothing step. How to choose the used smoothing parameter in a data-driven fashion remains an open and unsolved question. A last point we would mention is about the concentration inequality (Proposition 4.1) used in this paper. A restrictive assumption is made on the entries of the considered random matrices (boundedness of independent elements). Although it is natural to think about a truncation-like technique to get rid of this restriction, such results are lacking as far as we know.

Acknowledgements

The research of Jiaqi Chen was supported by the Chinese NSF grant 10571020 and Research Grant ARED 06007848 from Région Bretagne, France.

The authors thank the two referees for their detailed comments which have led to significant improvements of the manuscript.

References

[1] Z.D. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, Science Press, Beijing, 2006.
[2] Z.D. Bai, J.Q. Chen, J.F. Yao, On estimation of the population spectral distribution from a high-dimensional sample covariance matrix, Australian & New Zealand J. Statist. 52 (2010) 423–437.
[3] B. Delyon, Concentration inequalities for the spectral measure of random matrices, Electron. Commun. Probab. 15 (2010) 549–562.

[4] N. El Karoui, Spectrum estimation for large dimensional covariance matrices using random matrix theory, Ann. Statist. 36 (6) (2008) 2757–2790.
[5] I. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statistics 29 (2) (2001) 295–327.
[6] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. USSR-Sb 1 (1967) 457–483.
[7] N. Raj Rao, James A. Mingo, Roland Speicher, Alan Edelman, Statistical eigen-inference from large Wishart matrices, Ann. Statist. 36 (6) (2008) 2850–2885.
[8] Jack W. Silverstein, Sang-Il Choi, Analysis of the limiting spectral distribution of large-dimensional random matrices, J. Multivariate Anal. 54 (2) (1995) 295–309.
[9] G. Szegö, Orthogonal Polynomials, Americain Mathematical Society, New York, 1959.