On Generalized Expectation Based Estimation of a Population Spectral Distribution from High-Dimensional Data

Weiming Li[1] and Jianfeng Yao[2,3]
[1]Department of Mathematics, Beijing University of Posts and Telecommunications, Beijing 100876, CHINA
[2]Department of Statistics and Actuarial Sciences, The University of Hong Kong, Hongkong, CHINA
[3]Corresponding author: Jianfeng Yao, e-mail: jeffyao@hku.hk

Abstract

This paper discusses the problem of estimating the population spectral distribution from high-dimensional data. We present a general estimation procedure that covers situations where the moments of this distribution fail to identify the model parameters. The main idea is to use generalized functional expectations as a substitute for the moments. Beyond the consistency, we also prove a central limit theorem for the proposed estimator. An application to the analysis of the eigenvalues of the sample correlation matrix of S&P 500 daily stock returns is proposed.

Keywords: Large sample covariance matrix; Population spectral distribution; Empirical spectral distribution; Generalized expectation estimation.

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 $S_n = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^*/n$ 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.

More precisely, the *spectral distribution* (SD) F^A of an $m \times m$ Hermitian matrix (or real symmetric) A is the measure generated by its eigenvalues $\{\lambda_i^A\}$,

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

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 SD

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

This SD or its limit H (see below) is referred as the *population spectral distribution* (PSD) of the observation model.

The main observation is that for high-dimensional data, the observed SD $F_n = F^{S_n}$ of the sample covariance matrix is far from the PSD H_p. Indeed, under reasonable assumptions, when both dimensions p and n grow proportionally, almost

surely, the empirical SD F_n will weakly converge to a deterministic distribution F, called *limiting spectral distribution* (LSD), which in general has no explicit form but is expressed via an implicit equation (Marčenko and Pastur 1967; Silverstein 1995).

A natural question here is the recovering of the PSD H_p (or its limit H) from the sample covariance matrix S_n. This question has a central importance in several popular statistical methodologies like principal component analysis (Johnstone 2001) or factor analysis that all rely on an efficient estimation of some population covariance matrices.

Recent works on this problem include Mestre (2008) where the author introduces a method based on contour integration under an eigenvalue splitting condition. Most recently, Li and Yao (2013) has provided an extension of Mestre's method to situations where the eigenvalue splitting condition cannot be met. Some related references include El Karoui (2008), Rao et al. (2008), Bai et al. (2010), Chen et al. (2011), and Li et al. (2013).

However, except El Karoui (2008) and Li et al. (2013), all the cited estimation methods are based on the moments of the PSD H. It may happen and that has been a surprise, that these moments can not help to identify model parameters. Such an example is provided in Section 4, where the underlying PSD H has a normalized unit mean and infinite variance whatever the values of model parameter. Clearly, any estimation procedure based on the moments of H fails in such situations.

The main motivation of this work is to propose a new estimator to cover these intriguing situations. Inspired by the generalized method of moments, we consider empirical statistics linked to a class of general *test* functions f. These test functions are usually smaller than the monomials z^j and thus expected to have a finite expectation with respect to the unknown PSD H. In the example of stock returns data, H has a infinite variance but test functions like $f(x) = \sin(x)$ do have a finite integral with respect to H, which makes its estimation possible.

2. Generalized expectation estimation

Let G be a measure on the real line, the support of G is denoted by S_G. The Stieltjes transform of G is

$$s_G(z) = \int \frac{1}{x - z} dG(x), \quad z \in \mathbb{C}^+,$$

which is a one-to-one map defined on the upper half complex plane $\mathbb{C}^+ = \{z \in \mathbb{C} : \Im(z) > 0\}$. The transform can be trivially extended to $\mathbb{C} \setminus S_G$ by using the same functional form, which will be adopted throughout the paper.

Suppose that the underlying PSD H belongs to a parametric family $\mathcal{H} = \{H(\theta) : \theta \in \Theta \subset \mathbb{R}^q\}$. Denote by c the limiting ratio of p/n, and F the LSD with respect to H and c. Let f be an analytic function on an open region containing the support S_F of F, and $H(f)$ be the expectation of f with respect to H, i.e.

$$H(f) = \int f(t) dH(t).$$

We call this integral generalized expectation of the PSD H. It will be shown that $H(f)$ connects to F through the Stieltjes transform $\underline{s}(z)$ of $cF + (1 - c)\delta_0$ by a contour integral:

$$H(f) = K(c, f) + \frac{1}{2\pi \mathrm{i} c} \oint_C z\underline{s}'(z) f(-1/\underline{s}(z)) dz, \tag{2.1}$$

where $\underline{s}'(z)$ stands for the derivative of $\underline{s}(z)$, $K(c,f)$ is a constant related to c and f, and C is a positive oriented contour enclosing the support S_F (see Theorem 3.1). When an empirical SD $F_n := F^{S_n}$ is obtained, we may use the Stieltjes transform $\underline{s}_n(z)$ of $(p/n)F_n + (1-p/n)\delta_0$ and its derivative $\underline{s}_n'(z)$ to estimate $\underline{s}(z)$ and $\underline{s}'(z)$, respectively, in the formula (2.1), and then get an estimate

$$\widehat{H}(f) := K(p/n, f) + \frac{n}{p}\frac{1}{2\pi \mathrm{i}}\oint_C z\underline{s}_n'(z) f(-1/\underline{s}_n(z))dz. \tag{2.2}$$

Now with the help of $H(f)$ and its estimate $\widehat{H}(f)$, we consider the estimation of the PSD H. Let $f_1, \ldots, f_q$ be analytic functions on an open region containing S_F, $\boldsymbol{\gamma} = (H(f_j))_{1\le j\le q}$ be a q dimensional vector of generalized expectations. In order to make θ identifiable from $\boldsymbol{\gamma}$, we assume that the vector function g from $\mathbb{R}^q$ to $\mathbb{R}^q$: $\theta \mapsto \boldsymbol{\gamma}$ is invertible in Θ. Under this assumption, the *generalized expectation estimator* (GEE) of θ is

$$\widehat{\theta}_n = g^{-1}(\widehat{\boldsymbol{\gamma}}_n),$$

where $\widehat{\boldsymbol{\gamma}}_n = (\widehat{H}(f_j))_{1\le j\le q}$ with elements defined by (2.2).

3. Asymptotic properties

In this section, we study the asymptotic properties of the expectations $\{\widehat{H}(f_j)\}$ and the GEE $\widehat{\theta}_n$. All these properties are based on the following assumptions.

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}), $i, j \ge 1$ satisfying

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

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

Assumption (c). The PSD H_p of Σ_p weakly converges to a probability distribution H on $[0, \infty)$ as $n \to \infty$. Moreover, the sequence of spectral norms $(||\Sigma_p||)$ is bounded in p.

Assumptions (a)-(c) are classical conditions for the central limit theorem (CLT) of linear spectral statistics, see Bai and Silverstein (2004, 2010).

Theorem 3.1. *Under the assumptions* (a)-(c)*, for each* j $(1 \le j \le q)$*,*

(i) *the generalized expectation* $H(f_j)$ *can be re-expressed as*

$$H(f_j) = K(c, f_j) + \frac{1}{2\pi \mathrm{i} c}\oint_C z\underline{s}'(z) f_j(-1/\underline{s}(z))dz,$$

where C *is a positively oriented contour, taking values in* $\mathbb{C} \setminus (S_F \cup \{0\})$ *and enclosing the support* S_F *of* F*, and* $K(c, f_j) = (1 - 1/c) f_j(0)$ *if* C *enclosing 0, and zero otherwise*;

(ii) *the empirical expectation* $\widehat{H}(f_j)$ *based on* $\underline{s}_n(z)$ *converges almost surely.*

Theorem 3.2. *Under the assumptions* (a)-(c),

(i) *the random vector*

$$n\left(\widehat{H}(f_j) - H_p(f_j)\right)_{1\le j\le q} \tag{3.1}$$

forms a tight sequence in n*, where the centralization term* $H_p(f_j)$ *stands for the generalized expectation of* H_p*.*

(ii) *If* w_{11} *and* Σ_p *are real and* $\mathrm{E}(w_{11}^4) = 3$*, then* (3.1) *converges weakly to a Gaussian distribution* $N_q(\mu, \Phi)$*, with mean vector*

$$\mu = \left(-\frac{1}{2\pi \mathrm{i}} \oint_C f_j(-1/\underline{s}(z)) \frac{\int t^2 \underline{s}'(z)^2 dH(t)}{\underline{s}(z)(1+\underline{s}(z))^3} dz\right)_{1\le j\le q}$$

and covariance matrix $\Phi = (\phi_{ij})_{q\times q}$ *with*

$$\phi_{ij} = \frac{-1}{4\pi^2 c^2} \oint_C \oint_{C'} f_i(-1/\underline{s}(z_1)) f_j(-1/\underline{s}(z_2)) k(z_1, z_2) dz_1 dz_2,$$

where $k(z_1, z_2) = 2\underline{s}'(z_1)\underline{s}'(z_2)/(\underline{s}(z_1) - \underline{s}(z_2))^2 - 2/(z_1 - z_2)^2$*. The contours* C *and* C' *shares the same properties and are assumed non-overlapping.*

(iii) *If* w_{11} *is complex with* $\mathrm{E}(w_{11}^2) = 0$ *and* $\mathrm{E}(|w_{11}|^4) = 2$*, then* (ii) *also holds, except the mean vector is zero and the covariance matrix is* $\Phi/2$*.*

Next, we present the asymptotic properties of the GEE $\widehat{\theta}_n$.

Theorem 3.3. *In addition to the assumptions* (a)-(c)*, suppose that the true value of the parameter* θ_0 *is an inner point of* Θ*. Also, suppose that the function* $g(\theta)$ *is differentiable in a neighborhood of* θ_0 *and the Jacobian matrix* $J(\theta) = \partial g/\partial\theta$ *is invertible at* θ_0*. Then,*

(i) *the GEE* $\widehat{\theta}_n$ *is strongly consistent,*

(ii) *moreover, if assumptions in* (ii) *or* (iii) *of Theorem 3.2 on* w_{11} *hold, then*

$$n(\widehat{\theta}_n - g^{-1}(\boldsymbol{\gamma}_p)) \xrightarrow{\mathcal{D}} N_q(J^{-1}(\theta_0)\mu(\theta_0), \Gamma(\theta_0)),$$

where $\boldsymbol{\gamma}_p = (H_p(f_j))_{1\le j\le q}$ *and* $\Gamma(\theta_0) = J^{-1}(\theta_0)\Phi(\theta_0)(J^{-1}(\theta_0))'$*, with* μ *and* Φ *defined in Theorem 3.2.*

4. Application to S&P 500 daily stocks data

We consider an empirical correlation matrix of daily returns from stocks listed in the Standard & Poor Index and analyze the distribution of its eigenvalues. The time period is from September, 2007 to September 2011 covering 1001 trading days. As 12 stocks listed as by September 2011 do not have a complete history, they are removed from the analysis and in total 488 U.S. stocks have been included. The total data matrix of the returns is then with data dimension $p = 488$ and sample size

$n = 1000$. Next the 488×488 sample correlation matrix of these returns is computed and we obtain its 488 sample eigenvalues.

It is well known that for correlation matrices from stock returns or macroeconomic time series, a few large eigenvalues detach from the bulk of the eigenvalues and they are termed as *spikes*, see Johnstone (2001). Here we concentrate ourselves on the analysis of the bulk eigenvalues by removing the 6 first largest ones which are deemed as spike eigenvalues.

The question we address here is: what is the structure of the eigenvalues at the population level that has led to these observed eigenvalues. To this end and following Bouchaud and Potters (2009) and Li et al. (2013), an inverse cubic density is assumed for PSD H associated to the bulk eigenvalues, that is,

$$h(t|\alpha) = \frac{c}{(t-a)^3} I(t \geq \alpha), \quad 0 \leq \alpha < 1,$$

where $c = 2(1-\alpha)^2$ and $a = 2\alpha - 1$.

As already noticed, moments-based methods fail to estimate the parameter for α that the moments of $H(\alpha)$ can not identify the parameter: H has infinite variance and unit mean whatever the value of α. However, expectations of a suitably-chosen "test" function with respect to H can help to identify α.

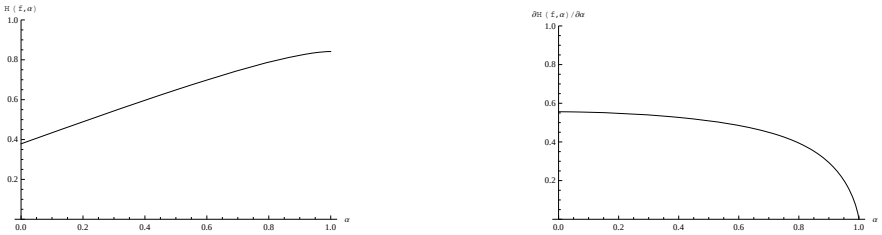

Figure 1: Curves of $H(f, \alpha)$ (left) and $\partial H(f, \alpha)/\partial \alpha$ (right).

Here, we provide an example with the test function $f(z) = \sin(z)$, that is, we consider the expectation

$$H(f, \alpha) = \int \sin(t) h(t|\alpha) dt,$$

which exists and is increasing with respect to α (although $H(f, \alpha)$ has no analytic expression), see Figure 1 .

The estimate of the expectation turns out to be $\widehat{H}(f, \alpha) = 0.5546$ which indicates $\widehat{\alpha} = 0.3205$. To make a comparison, we plot in Figure 2 the limiting spectral density predicted by the random matrix theory (derived from $H(\widehat{\alpha})$) and the empirical density function of bulk eigenvalues (an estimate using a Gaussian kernel with bandwidth $h = 0.01$). It shows that our estimation yields a very accurate fit to the empirical density. Potential applications in the future of these findings can be done through an explicit factor modeling where factor scores and loadings, once estimated, will provide important information on the correlations, at the population level, between returns of the listed stocks.

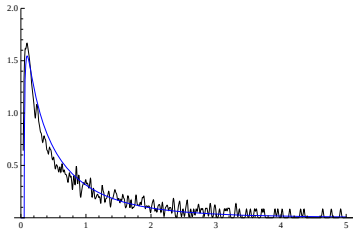

Figure 2: The empirical density of the sample eigenvalues (plain black) compared to the limiting spectral densities corresponding to $H(\widehat{\alpha})$ (blue).

References

Bai, Z. D., Chen, J. Q., Yao, J. F. (2010). On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.*, 52, 423–437.

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. (2010). *Spectral Analysis of Large Dimensional Random Matrices* (2nd ed.). New York: Springer.

Bouchaud, J. P., Potters, M. (2009). Financial applications of random matrix theory: A short review. ArXiv:09101205v1.

Chen, J. Q., Delyon, B., Yao, J. F. (2011). On a model selection problem from high-dimensional sample covariance matrices. *J. Multivariate Anal.*, 510, 1388–1398.

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

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

Li, W. M., Chen, J. Q., Qin, Y. L., Yao, J. F., Bai, Z. D. (2013). Estimation of the population spectral distribution from a large dimensional sample covariance matrix. ArXiv:1302.0355.

Li, W. M., Yao, J. F. (2013). A local moment estimation of the spectrum of a large dimensional covariance matrix. ArXiv:1302.0356.

Marčenko, V. A., Pastur, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72, 507–536.

Mestre, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Trans. Inform. Theory*, 54, 5113–5129.

Rao, N. R., 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.