Eigenvalues and eigenvectors of large sample covariance matrices

G.M. Pan
Eurandom, P.O.Box 513, 5600MB Eindhoven, the Netherlands.

Keywords: Statistics. Probability
AMS: 60J80

Abstract

This paper focuses on the theory of spectral analysis of Large sample covariance matrix. Concerning eigenvalues and eigenvectors some important results and methods are reviewed and moreover, some latest results are also presented.

1. Introduction

Random matrices theory (RMT) goes back as far as the quantum mechanics (QM) work in the 1940's and early 1950's. In QM, eigenvalues of Hermitian matrix can be used to describe the energy level of a system. Therefore the limiting behavior of large dimension RMT attracts those who work in QM. Since then the study of large dimension RMT has also appealed to considerable mathematicians, probabilitists and statisticians. Concerning the progress of RMT one may refer to the review of Bai (1999) and Methata (1990).

Moreover, much current research in statistics focuses on the problems posed by availability of large amounts of data such as genomics, finance and wireless communication. As a powerful tool, RMT provide a remarkable insight for such problems. For example, see Tulino and Verdu (2004).

In this paper we give a brief review of sample covariance matrices and some latest result are also presented.

2. Sample covariance matrices

Let $X_n = (X_{ij})$ be an $n \times N$ matrix of i.i.d. complex random variables and let T_n be an $n \times n$ nonnegative definite Hermitian matrix with a square root $T_n^{\frac{1}{2}}$. In this paper, we shall consider the matrix $A_n = \frac{1}{N} T_n^{\frac{1}{2}} X_n X_n^* T_n^{\frac{1}{2}}$. If T_n is nonrandom, then A_n can be considered as a sample covariance matrix of a sample drawn from a population with distribution of $T_n^{1/2}\mathbf{X}_{\cdot,1}$, where $\mathbf{X}_{\cdot,1} = (X_{11}, \cdots, X_{n1})'$. If T_n is an inverse of another sample covariance matrix, then the multivariate F matrix can be considered as a special case of the matrix A_n.

A fundamental object in RMT is the empirical spectral distribution function (ESD), given by

$$F^{A_n}(x) = \frac{1}{n} \sum_{i=1}^{n} I(\lambda_i \leq x), \tag{1}$$

where $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of A_n.

To investigate eigenvalues of A_n or (1), two popular methods have been used in RMT.

1. *Moment method.* A simple fact is that the kth moment of $F^{A_n}(x)$ is

$$\beta_k = \int x^k dF^{A_n}(x) = \frac{1}{n} trA_n^k.$$

Thus, to show that $F^{A_n}(x)$ converges to $F^c(x)$, a limiting distribution, one need to verify that

$$\frac{1}{n} trA_n^k \xrightarrow{a.s.} \int x^k dF^c(x).$$

and that Carleman's condition

$$\sum_{k=1}^{\infty} \beta_{2k}^{-1/2k} = \infty.$$

2. *Stietjes transform.* Another important mathematical tool in RMT is the **Stieltjes transform** which is defined by

$$m_G(z) = \int \frac{1}{\lambda - z} dG(\lambda), \quad z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C},\ \Im z > 0\}.$$

for any distribution function $G(x)$. It is well known that $G_n \xrightarrow{w} G$ if and only if $m_{G_n}(z) \to m_G(z)$, for all $z \in \mathbb{C}^+$.

M-P law When $T_n = I$, with probability one,

$$F^{A_n}(x) \to F^c(x), \tag{2}$$

whose density is

$$p_c(x) = \begin{cases} \frac{1}{2xc\pi}\sqrt{(b-x)(x-a)}, & \text{if } a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

and a point mass $1-1/c$ at the origin if $c > 1$ (see Marcenko and Pastur (1967) and Jonsson (1982)). Here $c = \lim_{n\to\infty} \frac{n}{N}$, $a = (1-\sqrt{c})^2$ and $b = (1+\sqrt{c})^2$. Later, for general A_n , Silverstein (1995) used the Stietjes transform to show that

$$F^{A_n}(x) \to F^{c,H}(x),$$

whose Stietjes transform, $m(z)$, is a solution to the equation

$$m(z) = \int \frac{1}{t(1-c-czm)-z} dH(t),$$

$H(t)$ the limiting spectral distribution of matrix T_n.

Extreme eigenvalues When $T_n = I$ and $0 < c < 1$, Bai and Yin (1993) used a unified approach to prove that

$$\lambda_{\max} \xrightarrow{a.s.} (1+\sqrt{c})^2, \quad \lambda_{\min} \xrightarrow{a.s.} (1-\sqrt{c})^2. \tag{3}$$

Further, Bai and Silverstein (1998, 1999) characterized the separation of eigenvalues of general A_n.

For the asymptotic distribution of $\lambda_{\max}$, if A_n is a Wishart matrix, Johnstone (2001) showed that

$$\frac{\lambda_{\max} - \mu_{nN}}{\sigma_{nN}} \xrightarrow{D} W_1 \sim F_1 \tag{4}$$

where

$$\mu_{nN} = (\sqrt{n-1} + \sqrt{N})^2, \quad \sigma_{nN} = (\sqrt{n-1} + \sqrt{N})(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{N}})^{1/3}$$

and F_1, the Tracy-Widom law of order 1, is given by

$$F_1(s) = exp\{-\frac{1}{2}\int_s^\infty q(x) + (x-s)q^2(x)dx\}$$

with $q(x)$ being the solution of Painleve II differential equation.

When $n = N$ and A_n is a subgaussian matrices, Rudelson and Vershynin (2007) established for any $\varepsilon \geq 0$

$$P(\lambda_{\min} \leq \varepsilon n^{-1/2}) \leq C\varepsilon + d^n \qquad 0 < d < 1 \tag{5}$$

and under more general matrices A_n Pan and Zhou (2007a) further gave

$$P(\lambda_{\min} \leq \varepsilon n^{-1/2}) \leq C\varepsilon + n^{-l}, \tag{6}$$

with l being any positive number.

central limit theorems of eigenvalues For the linear spectral statistic of eigenvalues the following central limit theorem was reported (Bai and Silverstein 2004). Let

$$L_N(x) = N(F^{\mathbf{A}_N}(x) - F^{c_N, H_N}(x)).$$

and $g_1, \cdots, g_k$ be analytic function, where $c_n = N/n$ and $F^{c_n, H_n}(x)$ is obtained from $F^{c,H}(x)$ with c and $H(x)$ replaced by c_n and H_n. Then under some assumptions

$$\left(\int g_1(x)dL_N(x), \cdots, \int g_k(x)dL_N(x)\right) \xrightarrow{D} (X_{g_1}, \cdots, X_{g_k}), \tag{7}$$

a gaussian vector with some mean and variance. Here one assumption needed is $Ev_{11}^4 = 3$ or $E|v_{11}|^4 = 2$. However this condition can be dropped when $\mathbf{T}_n = \mathbf{I}$ or is a diagonal matrix, as pointed out by Pan and Zhou (2007b).

Eigenvectors Regarding eigenvectors less work has been done except Silverstein (1989, 1990) and his earlier works.

Let $U_n \Lambda_n U_n^*$ denote the spectral decomposition of A_n, where $\Lambda_n = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$, $U_n = (u_{ij})$ is a unitary matrix consisting of the orthonormal eigenvectors of A_n. Assume that $\mathbf{x}_n \in \mathbb{C}^n, \|\mathbf{x}_n\| = 1$, is an arbitrary non-random unit vector and $\mathbf{y} = (y_1, y_2, \cdots, y_n)^* = U_n^* \mathbf{x}_n$.

We define a new empirical distribution function based on eigenvectors and eigenvalues as

$$F_1^{A_n}(x) = \sum_{i=1}^{n} |y_i|^2 I(\lambda_i \le x). \tag{8}$$

It is obvious that $F_1^{A_n}(x)$ is a random probability distribution function and its Stieltjes transform is given by

$$m_{F_1^{A_n}}(z) = x_n^*(A_n - zI)^{-1} x_n.$$

Under some assumptions Bai, Miao and Pan (2007) proved

$$F_1^{A_n}(x) \to F^{c,H}(x) \quad \text{a.s.}, \tag{9}$$

which indicates that $F_1^{A_n}(x)$ and $F^{A_n}(x)$ both have the identical asymptotical limit. Furthermore, if $f(x)$ is a bounded function, then

$$\sum_{j=1}^{n} |y_j^2| f(\lambda_j) - \frac{1}{n} \sum_{j=1}^{n} f(\lambda_j) \to 0 \quad \text{a.s.}.$$

To investigate central limit theorems of spectral statistic of eigenvectors and eigenvalues, let

$$G_n(x) = \sqrt{N}(F_1^{A_n}(x) - F^{c_n, H_n}(x)).$$

Then

$$\left(\int g_1(x) dG_n(x), \cdots, \int g_k(x) dG_n(x) \right) \xrightarrow{D} (X_{g_1}, \cdots, X_{g_k}) \tag{10}$$

a Gaussian vector with some mean and variance and $g_1, \cdots, g_k$ analytic functions (Bai, Miao and Pan (2007)). Similarly, the assumption $Ev_{11}^4 = 3$ or $E|v_{11}|^4 = 2$ is still needed in that paper. To remove it and guarantee central limit theorem, Pan and Zhou (2007b) imposed an addition condition, that is,

$$\max_i \left| \mathbf{e}_i^* \mathbf{T}_N^{1/2} (z \underline{m}(z) \mathbf{T}_N + zI)^{-1} x_N \right| \to 0, \tag{11}$$

where $\mathbf{e}_i$ is the $N \times 1$ column vector with the ith element being 1 and the rest being 0, which is implied by

$$\max_i |x_{Ni}| \to 0, \tag{12}$$

when T_n is a diagonal matrix.

3. Bibliography

[1] Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, A review. *Statistics Sinica* . **9** 611-677.

[2] Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of large dimensional sample covariance matrix. *Ann. Probab.* **21** 1275-1294.

[3] Bai, Z. D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large dimensional random matrices. *Ann. Probab.* **26** 316-345.

[4] Bai, Z. D. and Silverstein, J. W. (1999). Exact separation of eigenvalues of large dimensional sample convariance matrices. *Ann. Probab.* **27** 1536-1555.

[5] Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large dimensional sample covariance matrices. *Ann. Probab.* **32** 553-605.

[6] Bai, Z. D. and Miao, B. Q. and Pan, G. M. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.* **35** No.4

[7] Jonsson, D. (1982). Some limit theorems of the eigenvalues of sample convariance matrix. *J. Multivariate Anal.* **12** 1-38.

[8] Johnstone, I.M. (2001) On the distribution of the largest eigenvalue in principal component analysis. *Ann Statistics.* **2** 295-327.

[9] Marčenkon, V. A. and Pastur, L. A. (1967). Distribution for some sets of random matrices. *Math. USSR-Sb.* **1** 457-483.

[10] Rudelson, M. and Vershynin, R. (2007). The littlewood-offord problem and invertibility of random matrices, http://www.math.missouri.edu/~rudelson/papers/rv-invertibility.pdf.

[11] Pan, G.M. and Zhou, W. 2007(a). Circular law, Extreme Singular values and Potential theory, http://arxiv.org/abs/0705.3773.

[12] Pan, G.M. and Zhou, W. 2007(b). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. Submitted.

[13] Silverstein, J. W. (1995). Strong convergence of the limiting distribution of the eigenvalues of large dimensional random matrices. *J. Multivariate Anal.* **55** 331-339.

[14] Silverstein, J. W. (1990). Weak convergence of random functions defined by the eigenvectors of sample convariance matrices. *Ann. Probab.* **18** 1174-1193.

[15] Silverstein, J. W. (1989). On the eigenvectors of large dimensional sample covariance matrices. *J. Multivariate Anal.* **30** 1-16.