

A Note on Central Limit Theorems for Linear Spectral Statistics of Large Dimensional F -matrix

Shurong Zheng and Zhidong Bai

KLAS and School of Mathematics & Statistics, Northeast Normal University, P.R.China

E-mails: zhengsr@nenu.edu.cn; baizd@nenu.edu.cn

Abstract. Sample covariance matrix and multivariate F -matrix play important roles in multivariate statistical analysis. The central limit theorems (CLT) of linear spectral statistics associated with these matrices were established in Bai and Silverstein (2004) and Zheng (2012) which received considerable attentions and have been applied to solve many large dimensional statistical problems. However, the sample covariance matrices used in these papers are not centralized and there exist some questions about CLT's defined by the centralized sample covariance matrices. In this note, we shall provide some short complements on the CLT's in Bai and Silverstein (2004) and Zheng (2012), and show that the results in these two papers remain valid for the centralized sample covariance matrices, provided that the ratios of dimension p to sample sizes (n, n_1, n_2) are redefined as $p/(n - 1)$ and $p/(n_i - 1)$, $i = 1, 2$, respectively.

Main results

Let $\{X_{jk}, j, k = 1, 2, \dots\}$ and $\{Y_{jk}, j, k = 1, 2, \dots\}$ be two independent double arrays of independent random variables, either both real or both complex. In the sequel, we use A^* to denote a complex conjugate transpose of a vector or matrix \mathbf{A} . For $p > 1$, $n > 1$ and $N > 1$, we define $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ and $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_N)$ with column vectors $\mathbf{X}_j = (X_{j1}, \dots, X_{jp})'$, $1 \leq j \leq n$, and $\mathbf{Y}_k = (Y_{k1}, \dots, Y_{kp})'$, $1 \leq k \leq N$. Let \mathbf{T}_p be a $p \times p$ non-negative definite (nnd) matrix. There exists a unique nnd matrix $\mathbf{T}_p^{1/2}$ such that $\mathbf{T}_p = (\mathbf{T}_p^{1/2})^2$. Then, $(\mathbf{T}_p^{1/2}\mathbf{X}_1, \dots, \mathbf{T}_p^{1/2}\mathbf{X}_n)$ and $(\mathbf{T}_p^{1/2}\mathbf{Y}_1, \dots, \mathbf{T}_p^{1/2}\mathbf{Y}_N)$ can be considered as two independent samples of sizes n and N , respectively, drawn from a p -dimensional population with population covariance matrix \mathbf{T}_p . Let the *centralized* sample covariance matrix is

$$\mathbf{S}_x = \frac{1}{n-1} \left(\sum_{i=1}^n \mathbf{T}_p^{1/2} \mathbf{X}_i \mathbf{X}_i^* \mathbf{T}_p^{1/2} - n \mathbf{T}_p^{1/2} \bar{\mathbf{X}} \bar{\mathbf{X}}^* \mathbf{T}_p^{1/2} \right), \quad \mathbf{S}_y = \frac{1}{N-1} \left(\sum_{i=1}^N \mathbf{T}_p^{1/2} \mathbf{Y}_i \mathbf{Y}_i^* \mathbf{T}_p^{1/2} - N \mathbf{T}_p^{1/2} \bar{\mathbf{Y}} \bar{\mathbf{Y}}^* \mathbf{T}_p^{1/2} \right)$$

respectively, where $\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$ and $\bar{\mathbf{Y}} = \frac{1}{N} \sum_{i=1}^N \mathbf{Y}_i$ and the *centralized* F matrix is $\mathbf{F} = \mathbf{S}_x \mathbf{S}_y^{-1}$. The *simplified* sample covariance matrices are as $\mathbf{B}_x = \frac{1}{n} \sum_{i=1}^n \mathbf{T}_p^{1/2} \mathbf{X}_i \mathbf{X}_i^* \mathbf{T}_p^{1/2}$, $\mathbf{B}_y = \frac{1}{N} \sum_{i=1}^N \mathbf{T}_p^{1/2} \mathbf{Y}_i \mathbf{Y}_i^* \mathbf{T}_p^{1/2}$ and the *simplified* F-matrix is $\mathbf{G} = \mathbf{B}_x \mathbf{B}_y^{-1}$. It is mentioned that the *centralized* covariance matrix will have the same LSD as that of the corresponding *simplified* covariance matrix. We shall give the CLTs of linear spectral statistics (LSS) of *centralized* sample covariance matrix and *centralized* F matrix.

Theorem 0.1 *Assume that*

(a) *For each p , $\{X_{ij}, i \leq p, j \leq n\}$ are independent random variables with $EX_{ij} = 0$, $E|X_{11}|^2 = 1$, and satisfying*

$$\frac{1}{np} \sum_{j=1}^p \sum_{k=1}^n E|X_{jk}|^4 \mathbf{1}_{\{|X_{jk}| \geq \eta\sqrt{n}\}} \rightarrow 0, \quad \text{for any fixed } \eta > 0. \quad (0.1)$$

Note that the random variables may be allowed to depend on p , but we suppress this dependence from the notation for brevity.

(b) *We assume $E|X_{ij}|^4 = 3$ for the real case, and $E|X_{ij}|^4 = 2$ and $EX_{ij}^2 = 0$ for the complex case.*

(c) *$y_n = p/n \rightarrow y$, and*

(d) *\mathbf{T}_p is a $p \times p$ non-random nnd Hermitian matrix with bounded spectral norm in p , and its ESD $H_p \xrightarrow{D} H$ where H is a proper probability distribution.*

Let f be an analytic function on an open region in the complex plane which covers the support of LSD of \mathbf{S}_x with the origin excluded.

Then

(i) *the random variables*

$$X_p(f) = p \int f(x) d \left(F^{\mathbf{S}_x} - F^{\{y_{n-1}, H_p\}}(x) \right), \quad (0.2)$$

form a tight sequence in p , where $F^{\mathbf{S}_x}$ is the ESD of centralized sample covariance matrix \mathbf{S}_x , $F^{\{y, H\}}$ is the LSD of \mathbf{S}_x whose LSD's Stieltjes transform $m_y(z)$ satisfies $\underline{m}_y(z) = ym_y(z) - (1 - y)/z$ and $\underline{m}_y(z)$ is the unique solution to the equation

$$z = -\frac{1}{\underline{m}_y} + y \int \frac{t}{1 + t\underline{m}_y(z)} dH(t). \quad (0.3)$$

in the upper half complex plane for each $z \in \mathbb{C}^+ = \{z : \Im(z) > 0\}$.

(ii) *The random variables in (0.2) converges weakly to Gaussian variables X_f with the same means and covariance functions as given in Theorem 1.1 of Bai and Silverstein (2004).*

As for the CLT of LSS of \mathbf{F} matrix, we have the following theorem.

Theorem 0.2 Assume that

1. the two arrays $\{X_{jk}, j \leq p, k \leq n\}$ and $\{Y_{jk}, j \leq p, k \leq N\}$ satisfy for any fixed $\eta > 0$,

$$\frac{1}{np} \sum_{j=1}^p \sum_{k=1}^n E|X_{jk}|^4 \mathbf{1}_{\{|X_{jk}| \geq \eta\sqrt{n}\}} \rightarrow 0, \quad \frac{1}{Np} \sum_{j=1}^p \sum_{k=1}^N E|Y_{jk}|^4 \mathbf{1}_{\{|Y_{jk}| \geq \eta\sqrt{N}\}} \rightarrow 0. \quad (0.4)$$

2. For all j, k , $|EX_{jk}^4| = \beta_x + 1 + \kappa$, $|EY_{jk}^4| = \beta_y + 1 + \kappa$. If both \mathbf{X} and \mathbf{Y} are complex valued, then $EX_{jk}^2 = EY_{jk}^2 = 0$. Moreover, $y_n = p/n \rightarrow y_1 > 0$ and $y_N = p/N \rightarrow y_2 \in (0, 1)$.

Let f be an analytic function in an open region of the complex plane containing the interval $\left[\frac{(1-h)^2}{(1-y_2)^2}, \frac{(1+h)^2}{(1-y_2)^2} \right]$, the support of the continuous part of the LSD $F_{\mathbf{y}}$ of \mathbf{F} -matrix, $h = \sqrt{y_1 + y_2 - y_1 y_2}$ and $\mathbf{y} = (y_1, y_2)$.

Then, as $p \rightarrow \infty$, the random variables

$$W_p(f) = p \int f(x) d \left(F^{\mathbf{F}}(x) - F_{(y_{n-1}, y_{N-1})}(x) \right)$$

converges weakly to Gaussian variables $\{W_f\}$ which have the same means and covariance functions as given in Zheng (2012), where $F^{\mathbf{F}}(x)$ is the ESD of centralized F -matrix \mathbf{F} and $F_{(y_1, y_2)}(x)$ is the LSD defined by (2.4) of Zheng (2012).

References

Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.*, **32**(1A), 553-605.

Zheng, S. R. (2012). Central Limit Theorem for Linear Spectral Statistics of Large Dimensional F -Matrix. *Annales de l'Institut Henri Poincaré-Probabilités et Statistiques*, **48**(2), 444-476.