# The Convergence Of The Empirical Distribution Of Canonical Correlation Coefficients

Yanrong Yang, Guangming Pan *

**Abstract**

Suppose that $\{X_{jk}, j = 1, \cdots, p_1; k = 1, \cdots, n\}$ are independent and identically distributed (i.i.d) real random variables with $EX_{11} = 0$ and $EX_{11}^2 = 1$, and that $\{Y_{jk}, j = 1, \cdots, p_2; k = 1, \cdots, n\}$ are i.i.d real random variables with $EY_{11} = 0$ and $EY_{11}^2 = 1$, and that $\{X_{jk}, j = 1, \cdots, p_1; k = 1, \cdots, n\}$ are independent of $\{Y_{jk}, j = 1, \cdots, p_2; k = 1, \cdots, n\}$. This paper investigates the canonical correlation coefficients $r_1 \geq r_2 \geq \cdots \geq r_{p_1}$, whose squares $\lambda_1 = r_1^2, \lambda_2 = r_2^2, \cdots, \lambda_{p_1} = r_{p_1}^2$ are the eigenvalues of the matrix

$$\mathbf{S}_{xy} = \mathbf{A}_x^{-1}\mathbf{A}_{xy}\mathbf{A}_y^{-1}\mathbf{A}_{xy}^T,$$

where

$$\mathbf{A}_x = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^T,\ \mathbf{A}_y = \frac{1}{n}\sum_{k=1}^{n}\mathbf{y}_k\mathbf{y}_k^T,\ \mathbf{A}_{xy} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{y}_k^T,$$

and

$$\mathbf{x}_k = (X_{1k}, \cdots, X_{p_1k})^T,\ \mathbf{y}_k = (Y_{1k}, \cdots, Y_{p_2k})^T,\ k = 1, \cdots, n.$$

When $p_1 \to \infty$, $p_2 \to \infty$ and $n \to \infty$ with $\frac{p_1}{n} \to c_1$, $\frac{p_2}{n} \to c_2$, $c_1, c_2 \in (0, 1)$, it is proved that the empirical distribution of $r_1, r_2, \cdots, r_{p_1}$ converges, with probability one, to a fixed distribution under the finite second moment condition.

**Keywords:** Canonical correlation coefficients; Empirical spectral distribution; Random matrix; Stieltjes transform; Lindeberg's method.

*Y. R. Yang and G. M. Pan are with Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371.

# 1 Introduction

Canonical correlation analysis (CCA) deals with the relationship between two random variable sets. Suppose that there are two random variable sets: $\mathbf{x} = \{x_1, \ldots, x_{p_1}\}$, $\mathbf{y} = \{y_1, \ldots, y_{p_2}\}$, where $p_1 \le p_2$. Assume that there are $n$ observations for each of the $p_1 + p_2$ variables and they are grouped into $p_1 \times n$ random matrix $\mathbf{X} = (X_{ij})_{p_1 \times n}$ and $p_2 \times n$ random matrix $\mathbf{Y} = (Y_{ij})_{p_2 \times n}$ respectively. CCA seeks the linear combinations $\mathbf{a}^T\mathbf{x}$ and $\mathbf{c}^T\mathbf{y}$ that are most highly correlated, that is to maximize

$$r = Corr(\mathbf{a}^T\mathbf{x}, \mathbf{c}^T\mathbf{y}) = \frac{\mathbf{a}^T\Sigma_{\mathbf{xy}}\mathbf{c}}{\sqrt{\mathbf{a}^T\Sigma_{\mathbf{xx}}\mathbf{a}}\sqrt{\mathbf{c}^T\Sigma_{\mathbf{yy}}\mathbf{c}}}, \tag{1.1}$$

where $\Sigma_{\mathbf{xx}}$, $\Sigma_{\mathbf{yy}}$ are population covariance matrices for $\mathbf{x}$, $\mathbf{y}$ respectively; $\Sigma_{\mathbf{xy}}$ is the population covariance matrix between $\mathbf{x}$ and $\mathbf{y}$.

After finding the maximal correlation $r_1$ and associated combination vectors $\mathbf{a}_1$, $\mathbf{c}_1$, CCA considers seeking a second linear combination $\mathbf{a}_2^T\mathbf{x}$, $\mathbf{c}_2^T\mathbf{y}$ that has the maximal correlation among all linear combinations uncorrelated with $\mathbf{a}_1^T\mathbf{x}$, $\mathbf{c}_1^T\mathbf{y}$. This procedure can be iterated and successive canonical correlation coefficients $r_1, \ldots, r_{p_1}$ can be found. Substituting population covariance matrices with sample covariance matrices, $r_1, \ldots, r_{p_1}$ can be recast as the roots of the determinant equation

$$det(\mathbf{A}_{xy}\mathbf{A}_y^{-1}\mathbf{A}_{xy}^T - r^2\mathbf{A}_x) = 0, \tag{1.2}$$

where

$$\mathbf{A}_x = \frac{1}{n}\mathbf{X}\mathbf{X}^T,\ \mathbf{A}_y = \frac{1}{n}\mathbf{Y}\mathbf{Y}^T,\ \mathbf{A}_{xy} = \frac{1}{n}\mathbf{X}\mathbf{Y}^T.$$

About this point, one may refer to page 284 of Mardia, Kent and Bibby (1979). The roots of the determinant equation above go under many names, because they figure equally in discriminant analysis, canonical correlation analysis, and invariant tests of linear hypotheses in the multivariate analysis of variance. These are standard techniques in multivariate statistical analysis. Section 4 of Wachter (1980) described how to transform these statistical settings to the determinant equation form. Johnstone (2008) also gave its applications in these aspects in multivariate statistical analysis.

The empirical distribution of the canonical correlation coefficients $r_1, r_2, \cdots, r_{p_1}$ is defined as

$$F(x) = \frac{1}{p_1}\#\{i : r_i \le x\}, \tag{1.3}$$

where $\#\{\cdots\}$ denotes the cardinality of the set $\{\cdots\}$. When the two variable sets $\mathbf{x}$ and $\mathbf{y}$ are independent and each set consists of i.i.d Gaussian random variables, Wachter (1980) proved that the empirical distribution of $r_1, r_2, \cdots, r_{p_1}$ converges in probability and obtained an explicit expression for the limit of the empirical distribution when $p_1, p_2$ and $n$ are all approaching infinity. From the determinant equation (1.2), it can be seen that $\lambda_1 = r_1^2, \lambda_2 = r_2^2, \ldots, \lambda_{p_1} = r_{p_1}^2$ are eigenvalues of the matrix $\mathbf{S}_{xy} = \mathbf{A}_x^{-1}\mathbf{A}_{xy}\mathbf{A}_y^{-1}\mathbf{A}_{xy}^T$. Hence the analysis of the empirical distribution of $r_1, r_2, \cdots, r_{p_1}$ is equivalent to analyzing the ESD of the matrix $\mathbf{S}_{xy}$. Here for any $p \times p$ matrix $\mathbf{A}$ with real eigenvalues $x_1 \le x_2 \le \ldots \le x_p$, its ESD is defined as

$$F^{\mathbf{A}}(x) = \frac{1}{p}\#\{i : x_i \le x\}. \tag{1.4}$$

The aim of this paper is to prove that the result in Wachter (1980) remains true when the entries of $\mathbf{X}$ and $\mathbf{Y}$ have finite second moments but not necessarily Gaussian distribution.

**Theorem 1.** *Assume that*
*(a)* $\mathbf{X} = (X_{ij})_{1\le i\le p_1, 1\le j\le n}$ *where* $X_{ij}, 1 \le i \le p_1, 1 \le j \le n$, *are i.i.d real random variables with* $EX_{11} = 0$ *and* $E|X_{11}|^2 = 1$.
*(b)* $\mathbf{Y} = (Y_{ij})_{1\le i\le p_2, 1\le j\le n}$ *where* $Y_{ij}, 1 \le i \le p_2, 1 \le j \le n$ *are i.i.d real random variables with* $EY_{11} = 0$ *and* $E|Y_{11}|^2 = 1$.
*(c)* $p_1 = p_1(n)$ *and* $p_2 = p_2(n)$ *with* $\frac{p_1}{n} \to c_1$ *and* $\frac{p_2}{n} \to c_2$, $c_1, c_2 \in (0,1)$, *as* $n \to \infty$.
*(d)* $\mathbf{S}_{xy} = \mathbf{A}_x^{-1}\mathbf{A}_{xy}\mathbf{A}_y^{-1}\mathbf{A}_{xy}^T$ *where* $\mathbf{A}_x = \frac{1}{n}\mathbf{X}\mathbf{X}^T$, $\mathbf{A}_y = \frac{1}{n}\mathbf{Y}\mathbf{Y}^T$ *and* $\mathbf{A}_{xy} = \frac{1}{n}\mathbf{X}\mathbf{Y}^T$.
*(e)* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent.*

*Then as* $n \to \infty$ *the empirical distribution of the matrix* $r_1, r_2, \cdots, r_{p_1}$ *converges almost surely to a fixed distribution function whose density is*

$$\rho(r) = ((r-L)(r+L)(H-r)(H+r))^{\frac{1}{2}}/[\pi c_1 r(1-r)(1+r)], \ r \in [L, H], \tag{1.5}$$

*where* $L = |(c_2 - c_2c_1)^{\frac{1}{2}} - (c_1 - c_1c_2)^{\frac{1}{2}}|$ *and* $H = |(c_2 - c_2c_1)^{\frac{1}{2}} + (c_1 - c_1c_2)^{\frac{1}{2}}|$; *and atoms of size* $max(0, 1 - c_2/c_1)$ *at zero and size* $max(0, 1-(1-c_2)/c_1)$ *at unity.*

**Remark 1.** *The inverse of a matrix, such as* $\mathbf{A}_x^{-1}$ *and* $\mathbf{A}_y^{-1}$, *is the Moore-Penrose pseudoinverse, i.e. in the spectral decomposition of the initial matrix, replace each nonzero eigenvalue by its reciprocal and leave the zero eigenvalues alone. This is because under the finite second moment condition, the matrices* $\mathbf{A}_x$ *and* $\mathbf{A}_y$ *may be not invertible under the classical inverse matrix definition. However, with the additional assumption that* $EX_{11}^4 < \infty$ *and* $EY_{11}^4 < \infty$, *we have the conclusion that the smallest eigenvalues of the sample matrices* $\mathbf{A}_x$ *and* $\mathbf{A}_y$ *converge to* $(1-\sqrt{c_1})^2$ *and* $(1-\sqrt{c_2})^2$ *respectively[Theorem 5.11 of Bai and Silverstein (2009)], which are not zero since* $c_1, c_2 \in (0,1)$. *So* $\mathbf{A}_x$ *and* $\mathbf{A}_y$ *are invertible with probability one under the finite fourth moment condition.*

As stated previously, it is sufficient to analyze the limiting spectral distribution (LSD) of the matrix $\mathbf{S}_{xy}$, where LSD denotes the limit of the empirical spectral distribution as $n \to \infty$.

The strategy of the proof of Theorem 1 is as follows. Since the matrix $\mathbf{S}_{xy}$ is not symmetric, it is difficult to work on it directly. Instead we consider the $n \times n$ symmetric matrix

$$\mathbf{P}_y\mathbf{P}_x\mathbf{P}_y \tag{1.6}$$

where

$$\mathbf{P}_x = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}, \ \mathbf{P}_y = \mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}.$$

Note that $\mathbf{P}_x$ and $\mathbf{P}_y$ are projection matrices. It is easy to see that the eigenvalues of the matrix $\mathbf{P}_y\mathbf{P}_x\mathbf{P}_y$ are the same as those of the matrix $\mathbf{S}_{xy}$ other than $n-p_1$ zero eigenvalues, i.e.

$$F^{\mathbf{P}_y\mathbf{P}_x\mathbf{P}_y}(x) = \frac{p_1}{n}F^{\mathbf{S}_{xy}}(x) + \frac{n-p_1}{n}I_{[0,+\infty)}(x). \tag{1.7}$$

By (1.7) and the result in Wachter (1980), one can easily obtain the limit of $F^{\mathbf{P}_y\mathbf{P}_x\mathbf{P}_y}(x)$ when the entries of $\mathbf{X}$ and $\mathbf{Y}$ are Gaussian distributed. To move from the Gaussian case to non-Gaussian case, we mainly use Lindeberg's method (see Lindeberg (1922) and Chatterjee (2006)) and the Stieltjes transform. The Stieltjes transform for any probability distribution function $G(x)$ is defined as

$$m_G(z) = \int \frac{1}{x-z}dG(x), \quad z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C},\ v = \Im z > 0\}. \tag{1.8}$$

An additional key technique is to introduce a perturbation matrix in order to deal with the random matrix $(\mathbf{X}\mathbf{X}^T)^{-1}$ under the finite second moment condition.

# 2 Proof of Theorem 1

We divide the proof of Theorem 1 into 4 parts:

## 2.1 Step 1: Introducing a perturbation matrix

Let

$$\mathbf{A} = \mathbf{P}_y\mathbf{P}_x\mathbf{P}_y.$$

In view of (1.7) it is enough to investigate $F^{\mathbf{A}}$ to prove Theorem 1. In order to deal with the matrix $(\mathbf{X}\mathbf{X}^T)^{-1}$, we make a perturbation of the matrix $\mathbf{A}$ and obtain a new matrix

$$\mathbf{B} = \mathbf{P}_y\mathbf{P}_{tx}\mathbf{P}_y,$$

where $\mathbf{P}_{tx} = \frac{1}{n}\mathbf{X}^T(\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-1}\mathbf{X}$, $t > 0$ is a small constant number and $\mathbf{I}_{p_1}$ is the identity matrix of the size $p_1$.
We claim that, with probability one,

$$\lim_{t\to 0}\lim_{n\to\infty} L\Big(F^{\mathbf{A}}, F^{\mathbf{B}}\Big) = 0. \tag{2.1}$$

where $L(F^{\mathbf{A}}, F^{\mathbf{B}})$ is the Levy distance between two distribution functions $F^{\mathbf{A}}(\lambda)$ and $F^{\mathbf{B}}(\lambda)$. By Lemma 6 in the Appendix,

$$\begin{aligned} L^3(F^{\mathbf{A}}, F^{\mathbf{B}}) &\le \frac{1}{n}tr(\mathbf{A}-\mathbf{B})^2 \le \frac{1}{n}tr(\mathbf{P}_x - \mathbf{P}_{tx})^2 \\ &= \frac{1}{n}tr\big(\frac{1}{n}\mathbf{X}\mathbf{X}^T[(\frac{1}{n}\mathbf{X}\mathbf{X}^T)^{-1} - (\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-1}]\big)^2 \\ &\le \frac{t^2}{n}tr(\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-2}, \end{aligned} \tag{2.2}$$

where the second inequality uses the fact that $||\mathbf{P}_y|| = 1$ with the norm being the spectral norm and the last inequality uses the definition of the Moore-Penrose pseudoinverse so that we may write

$$
\begin{aligned}
& \frac{1}{n}\mathbf{X}\mathbf{X}^T[(\frac{1}{n}\mathbf{X}\mathbf{X}^T)^{-1} - (\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-1}] \\
= & \ \mathbf{U}^T \begin{pmatrix} \mu_1 & & & & & \\ & \ddots & & & & \\ & & \mu_m & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \mathbf{U}\mathbf{U}^T \begin{pmatrix} \frac{t}{\mu_1(\mu_1+t)} & & & & & \\ & \ddots & & & & \\ & & \frac{t}{\mu_m(\mu_m+t)} & & & \\ & & & -\frac{1}{t} & & \\ & & & & \ddots & \\ & & & & & -\frac{1}{t} \end{pmatrix} \mathbf{U} \\
= & \ \mathbf{U}^T \begin{pmatrix} \frac{t}{\mu_1+t} & & & & & \\ & \ddots & & & & \\ & & \frac{t}{\mu_m+t} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \mathbf{U}.
\end{aligned}
$$

Here $\mu_1, \ldots, \mu_m$ are the nonzero eigenvalues of the matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^T$ and $\mathbf{U}^T$ is the eigenvectors matrix of $\frac{1}{n}\mathbf{X}\mathbf{X}^T$.

Given $t > 0$, by Theorem 3.6 in Bai and Silverstein (2009) (or see Jonsson (1982) and Marčenko and Pastur (1967)) and the Helly-Bray theorem, we have with probability one

$$
\begin{aligned}
\frac{1}{n}tr(\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-2} &= \frac{p_1}{n}\int \frac{1}{(\lambda+t)^2} dF_{p_1}(\lambda) \to c_1 \int_a^b \frac{1}{(\lambda+t)^2} dF_{c_1}(\lambda) \\
&= \int_a^b \frac{\sqrt{(b-\lambda)(\lambda-a)}}{(\lambda+t)^2 2\pi\lambda} d\lambda \le \int_a^b \frac{\sqrt{(b-\lambda)(\lambda-a)}}{\lambda^3 2\pi} d\lambda \le M,
\end{aligned}
$$

where $F_{p_1}$ is the ESD of the sample matrix $\frac{1}{n}\mathbf{X}\mathbf{X}^T$, $F_{c_1}$ is the Marcenko-Pastur Law, $b = (1+\sqrt{c_1})^2$ and $a = (1-\sqrt{c_1})^2$. Here and in what follows $M$ stands for a positive constant number and it may be different from line to line. This, together with (2.2), implies (2.1), as claimed.

Let $\bar{\mathbf{B}}$ and $\bar{\mathbf{A}}$, respectively, denote analogues of the matrices $\mathbf{B}$ and $\mathbf{A}$ with the elements of $\mathbf{X}$ replaced by i.i.d. Gaussian distributed random variables, independent of the entries of $\mathbf{Y}$. By (2.1) and the fact that, for any $\lambda \in \mathbb{R}$,

$$
|F^{\mathbf{A}}(\lambda) - F^{\bar{\mathbf{A}}}(\lambda)| \le |F^{\mathbf{A}}(\lambda) - F^{\mathbf{B}}(\lambda)| + |F^{\mathbf{B}}(\lambda) - F^{\bar{\mathbf{B}}}(\lambda)| + |F^{\bar{\mathbf{B}}}(\lambda) - F^{\bar{\mathbf{A}}}(\lambda)|,
$$

in order to prove that, for any fixed $t > 0$, with probability one,

$$
\lim_{n\to\infty} |F^{\mathbf{A}}(\lambda) - F^{\bar{\mathbf{A}}}(\lambda)| = 0, \tag{2.3}
$$

it suffices to prove with probability one,

$$\lim_{n\to\infty} |F^{\mathbf{B}}(\lambda) - F^{\bar{\mathbf{B}}}(\lambda)| = 0. \tag{2.4}$$

If we have (2.3), then for any $\lambda \in \mathbb{R}$, with probability one,

$$\lim_{n\to\infty} |F^{\mathbf{P}_x\mathbf{P}_y}(\lambda) - F^{\mathbf{P}_x^g\mathbf{P}_y}(\lambda)| = 0. \tag{2.5}$$

Since $\mathbf{P}_y$ and $\mathbf{P}_x$ stand symmetric positions in the matrix $\mathbf{P}_x\mathbf{P}_y$, as in (2.3) and (2.5), one can similarly prove that for any $\lambda \in \mathbb{R}$, with probability one,

$$\lim_{n\to\infty} |F^{\mathbf{P}_x^g\mathbf{P}_y}(\lambda) - F^{\mathbf{P}_x^g\mathbf{P}_y^g}(\lambda)| = 0, \tag{2.6}$$

where $\mathbf{P}_y^g$ is obtained from the matrix $\mathbf{P}_y$ with all the entries of $\mathbf{Y}$ replaced by i.i.d Gaussian distributed random variables, independent of $\mathbf{P}_x^g$. Then (2.5) and (2.6) imply that for any $\lambda \in \mathbb{R}$, with probability one,

$$\lim_{n\to\infty} |F^{\mathbf{P}_x\mathbf{P}_y}(\lambda) - F^{\mathbf{P}_x^g\mathbf{P}_y^g}(\lambda)| = 0. \tag{2.7}$$

With the theorem obtained in Wachter (1980) and (2.7), our theorem is easily derived.

Hence the subsequent parts are devoted to proving (2.4).

## 2.2 Step 2: Truncation, Centralization, Rescaling and Tightness of $F^{\mathbf{B}}$

With (1.8) of Bai and Silverstein (2004) and the arguments above and below, we can choose $\varepsilon_n > 0$ such that $\varepsilon_n \to 0$, $n^{1/2}\varepsilon_n \to \infty$ as $n \to \infty$, and $P(|X_{ij}| \geq n^{1/2}\varepsilon_n) \leq \frac{\varepsilon_n}{n}$. Define

$$\tilde{X}_{ij} = X_{ij}I(|X_{ij}| < n^{1/2}\varepsilon_n),\ \hat{X}_{ij} = \tilde{X}_{ij} - E\tilde{X}_{11},$$
$$\mathbf{P}_{tx} = \frac{1}{n}\mathbf{X}^T(\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-1}\mathbf{X},\ \tilde{\mathbf{P}}_{tx} = \frac{1}{n}\tilde{\mathbf{X}}^T(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}\tilde{\mathbf{X}},$$
$$\hat{\mathbf{P}}_{tx} = \frac{1}{n}\hat{\mathbf{X}}^T(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}\hat{\mathbf{X}},\ \tilde{\mathbf{B}} = \mathbf{P}_y\tilde{\mathbf{P}}_{tx}\mathbf{P}_y,\ \hat{\mathbf{B}} = \mathbf{P}_y\hat{\mathbf{P}}_{tx}\mathbf{P}_y,$$

where $\tilde{\mathbf{X}} = (\tilde{X}_{ij})_{1\leq i\leq p_1;1\leq j\leq n}$ and $\hat{\mathbf{X}} = (\hat{X}_{ij})_{1\leq i\leq p_1;1\leq j\leq n}$.

Let $\eta_{ij} = 1 - I(|X_{ij}| < n^{1/2}\varepsilon_n)$. We then get by Lemma 4 in the appendix

$$\begin{aligned} &\sup_\lambda |F^{\mathbf{B}}(\lambda) - F^{\tilde{\mathbf{B}}}(\lambda)| \leq \frac{1}{n}rank(\mathbf{P}_y\mathbf{P}_{tx}\mathbf{P}_y - \mathbf{P}_y\tilde{\mathbf{P}}_{tx}\mathbf{P}_y) \leq \frac{1}{n}rank(\mathbf{P}_{tx} - \tilde{\mathbf{P}}_{tx}) \\ \leq\ & \frac{1}{n}[rank(\mathbf{X}^T - \tilde{\mathbf{X}}) + rank(\mathbf{X}\mathbf{X}^T - \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T) + rank(\mathbf{X} - \tilde{\mathbf{X}}^T)] \leq \frac{4}{n}\sum_{i=1}^{p_1}\sum_{j=1}^{n}\eta_{ij}. \end{aligned}$$

Denote $q = P(\eta_{ij} = 1) = P(|X_{ij}| \geq n^{1/2}\varepsilon_n)$. We conclude from Lemma 5 that for any $\delta > 0$,

$$P(\sup_\lambda |F^{\mathbf{B}}(\lambda) - F^{\tilde{\mathbf{B}}}(\lambda)| \geq \delta) \leq P(\frac{1}{n}\sum_{i=1}^{p_1}\sum_{j=1}^{n}\eta_{ij} \geq \delta)$$

$$= P(\sum_{i=1}^{p_1}\sum_{j=1}^{n}\eta_{ij} - np_1q \geq np_1(\frac{\delta}{p_1} - q))$$

$$\leq 2exp\Big( - \frac{n^2p_1^2(\frac{\delta}{p_1} - q)^2}{2np_1q + np_1(\frac{\delta}{p_1} - q)}\Big) \leq 2exp(-nh),$$

for some positive $h$. It follows from Borel-Cantelli's lemma that

$$\sup_{\lambda}|F^{\mathbf{B}}(\lambda) - F^{\tilde{\mathbf{B}}}(\lambda)| \to 0, \quad a.s. \quad \text{as } n \to \infty.$$

Next, we prove that

$$\sup_{\lambda}|F^{\hat{\mathbf{B}}}(\lambda) - F^{\tilde{\mathbf{B}}}(\lambda)| \to 0, \quad a.s. \quad \text{as } n \to \infty. \tag{2.8}$$

Again by Lemma 4 we have

$$\sup_{\lambda}|F^{\hat{\mathbf{B}}}(\lambda) - F^{\tilde{\mathbf{B}}}(\lambda)| \leq \frac{1}{n}rank(\hat{\mathbf{B}} - \tilde{\mathbf{B}}) \leq \frac{1}{n}rank\Big[\hat{\mathbf{P}}_{tx} - \tilde{\mathbf{P}}_{tx}\Big]$$

$$\leq \frac{1}{n}rank\Big[\frac{1}{n}\tilde{\mathbf{X}}^T\Big((\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1} - (\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}\Big)\tilde{\mathbf{X}}\Big]$$
$$+\frac{1}{n}rank\Big[\frac{1}{n}\tilde{\mathbf{X}}^T(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}E\tilde{\mathbf{X}}\Big] + \frac{1}{n}rank\Big[\frac{1}{n}(E\tilde{\mathbf{X}}^T)(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}\tilde{\mathbf{X}}\Big]$$
$$+\frac{1}{n}rank\Big[\frac{1}{n}(E\tilde{\mathbf{X}}^T)(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}E\tilde{\mathbf{X}}\Big].$$

Since all elements of $E\tilde{\mathbf{X}}$ are identical, $rank\Big(E\tilde{\mathbf{X}}\Big) = 1$. Moreover, from (2.10)

$$\begin{aligned}
& (\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1} - (\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1} \\
= \; & (\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T - \frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T)(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1} \\
= \; & \frac{1}{n}(\frac{1}{n}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}(-E\tilde{\mathbf{X}}E\tilde{\mathbf{X}}^T + \tilde{\mathbf{X}}E\tilde{\mathbf{X}}^T + (E\tilde{\mathbf{X}})\tilde{\mathbf{X}}^T)(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}.
\end{aligned}$$

Hence

$$\sup_{\lambda}|F^{\hat{\mathbf{B}}}(\lambda) - F^{\tilde{\mathbf{B}}}(\lambda)| \leq \frac{M}{n} \to 0.$$

Let $\hat{\sigma}^2 = E(|\hat{X}_{ij}|^2)$ and $\hat{\hat{\mathbf{B}}} = \frac{1}{n\hat{\sigma}^2}\hat{\mathbf{X}}^T(\frac{1}{n\hat{\sigma}^2}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}\hat{\mathbf{X}}$. Then by Lemma 6, we have

$$\begin{aligned}
& L^3(F^{\hat{\mathbf{B}}}, F^{\hat{\hat{\mathbf{B}}}}) \leq \frac{1}{n} tr(\hat{\mathbf{B}} - \hat{\hat{\mathbf{B}}})^2 \\
= \ & \frac{(\hat{\sigma}^2-1)^2t^2}{n} tr\Big(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \hat{\sigma}^2t\mathbf{I}_{p_1})^{-1}(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}\Big)^2 \\
= \ & \frac{(\hat{\sigma}^2-1)^2t^2}{n} tr\Big((\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \hat{\sigma}^2t\mathbf{I}_{p_1} - \hat{\sigma}^2t\mathbf{I}_{p_1})(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \hat{\sigma}^2t\mathbf{I}_{p_1})^{-1}(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}\Big)^2 \\
= \ & \frac{(\hat{\sigma}^2-1)^2t^2}{n} tr\Big((\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1} - \hat{\sigma}^2t(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \hat{\sigma}^2t\mathbf{I}_{p_1})^{-1}(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}\Big)^2 \\
\leq \ & \frac{(\hat{\sigma}^2-1)^2t^2}{n} p_1\Big(||(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}|| + \hat{\sigma}^2t||(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \hat{\sigma}^2t\mathbf{I}_{p_1})^{-1}|| \cdot ||(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}||\Big)^2 \\
\leq \ & \frac{(\hat{\sigma}^2-1)^2t^2}{n} p_1 \frac{4}{t^2} \to 0,
\end{aligned}$$

because $\hat{\sigma}^2 \to 1$ and $p_1/n \to c_1$ as $n \to \infty$; where the first equality uses the formula (2.10); the second inequality uses the matrix inequality that

$$tr(\mathbf{C}) \leq p_1||\mathbf{C}||,$$

holding for any $p_1 \times p_1$ normal matrix $\mathbf{C}$; and the last inequality uses the fact that

$$||(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + \hat{\sigma}^2t\mathbf{I}_{p_1})^{-1}|| \leq \frac{1}{\hat{\sigma}^2t}, \ ||(\frac{1}{n}\hat{\mathbf{X}}\hat{\mathbf{X}}^T + t\mathbf{I}_{p_1})^{-1}|| \leq \frac{1}{t}.$$

In view of the truncation, centralization and rescaling steps above, in the sequel, we shall assume that the underlying variables satisfy

$$|X_{ij}| \leq n^{1/2}\varepsilon_n, \quad EX_{ij} = 0, \quad EX_{ij}^2 = 1, \tag{2.9}$$

and for simplicity we shall still use notation $X_{ij}$ instead of $\hat{X}_{ij}$.

We now turn to investigating the tightness of $F^{\mathbf{B}}$. For any constant number $K > 0$,

$$\int_{\lambda > K} dF^{\mathbf{B}} \leq \frac{1}{K}\int \lambda dF^{\mathbf{B}} = \frac{1}{K}\frac{1}{n} tr[\mathbf{P}_y\mathbf{P}_{tx}\mathbf{P}_y]$$

Since the largest eigenvalue of $\mathbf{P}_y$ is 1 and $\mathbf{P}_{tx}$ is a nonnegative matrix we obtain

$$\begin{aligned}
tr[\mathbf{P}_y\mathbf{P}_{tx}\mathbf{P}_y] &= tr[\mathbf{P}_y\mathbf{P}_{tx}] \\
&\leq tr[\mathbf{P}_{tx}] = tr[\frac{1}{n}\mathbf{X}\mathbf{X}^T(\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-1}] \leq n.
\end{aligned}$$

The last inequality has used the facts that $t > 0$ and that all the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^T(\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-1}$ are less than 1.

It follows that $F^{\mathbf{B}}$ is tight.

## 2.3 Step 3: Convergence of the random part

The aim in this section is to prove that

$$\frac{1}{n}tr\mathbf{B}^{-1}(z) - E\frac{1}{n}tr\mathbf{B}^{-1}(z) \to 0 \quad \text{a.s. as } n \to \infty.$$

To this end we introduce some notation. Let $\mathbf{x}_k$ denote the $k$th column of $\mathbf{X}$ and $\mathbf{e}_k$ the column vector of the size of $p_1$ with the $k$th element being 1 and otherwise 0. Moreover, define $\mathbf{X}_k$ to be the matrix obtained from $\mathbf{X}$ by replacing the elements of the $k$th column of $\mathbf{X}$ with 0.

Fix $v = \Im z > 0$. Define $\mathcal{F}_k$ to be the $\sigma$-field generated by $\mathbf{x}_1, \cdots, \mathbf{x}_k$. Let $E_k(\cdot)$ denote the conditional expectation with respect to $\mathcal{F}_k$ and $E_0$ denote expectation. That is, $E_k(\cdot) = E(\cdot|\mathcal{F}_k)$ and $E_0(\cdot) = E(\cdot)$. Let

$$\mathbf{B}^{-1}(z) = (\mathbf{P}_y\mathbf{P}_{tx}\mathbf{P}_y - z\mathbf{I})^{-1},\ \mathbf{B}_k = \mathbf{P}_y\mathbf{P}_k^{tx}\mathbf{P}_y,\ \mathbf{B}_k^{-1}(z) = (\mathbf{P}_y\mathbf{P}_k^{tx}\mathbf{P}_y - z\mathbf{I})^{-1},$$

where $\mathbf{P}_{tx} = \frac{1}{n}\mathbf{X}^T(\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-1}\mathbf{X}$, $\mathbf{P}_k^{tx} = \frac{1}{n}\mathbf{X}_k^T(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^T + t\mathbf{I}_{p_1})^{-1}\mathbf{X}_k$.
Define $\mathbf{H}_k^{-1} = (\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^T + t\mathbf{I}_{p_1})^{-1}$ and $\mathbf{H}^{-1} = (\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1})^{-1}$.

Note that $\mathbf{X} = \mathbf{X}_k + \mathbf{x}_k\mathbf{e}_k^T$, that the elements of $\mathbf{X}_k\mathbf{e}_k$ are all zero and hence that

$$\mathbf{X}\mathbf{X}^T - \mathbf{X}_k\mathbf{X}_k^T = \mathbf{x}_k\mathbf{x}_k^T.$$

This implies that

$$\mathbf{H}_k^{-1} - \mathbf{H}^{-1} = \frac{1}{n}\mathbf{H}^{-1}\mathbf{x}_k\mathbf{x}_k^T\mathbf{H}_k^{-1} = \frac{1}{1 + \frac{1}{n}\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k}\frac{1}{n}\mathbf{H}_k^{-1}\mathbf{x}_k\mathbf{x}_k^T\mathbf{H}_k^{-1},$$

where we make use of the formula

$$\mathbf{A}_1^{-1} - \mathbf{A}_2^{-1} = \mathbf{A}_2^{-1}(\mathbf{A}_2 - \mathbf{A}_1)\mathbf{A}_1^{-1}, \tag{2.10}$$

holding for any two invertible matrices $\mathbf{A}_1$ and $\mathbf{A}_2$;
and

$$(\mathbf{U} + \mathbf{u}\mathbf{v}^T)^{-1}\mathbf{u} = \frac{\mathbf{U}^{-1}\mathbf{u}}{1 + \mathbf{v}^T\mathbf{U}^{-1}\mathbf{u}}, \tag{2.11}$$

holding for any invertible matrices $\mathbf{U}$ and $(\mathbf{U} + \mathbf{u}\mathbf{v}^T)$, vectors $\mathbf{u}$ and $\mathbf{v}$. We then write

$$\mathbf{B}_k - \mathbf{B} = \mathbf{P}_y(\mathbf{P}_k^{tx} - \mathbf{P}_{tx})\mathbf{P}_y = \mathbf{P}_y(C_1 + C_2 + C_3 + C_4)\mathbf{P}_y, \tag{2.12}$$

where

$$C_1 = \frac{1}{n}\frac{\mathbf{X}_k^T\mathbf{H}_k^{-1}\frac{1}{n}\mathbf{x}_k\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{X}_k}{1 + \frac{1}{n}\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k}, \quad C_2 = -\frac{1}{n}\frac{\mathbf{X}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k\mathbf{e}_k^T}{1 + \frac{1}{n}\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k},$$
$$C_3 = -\frac{1}{n}\frac{\mathbf{e}_k\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{X}_k}{1 + \frac{1}{n}\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k}, \quad C_4 = -\frac{1}{n}\frac{\mathbf{e}_k\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k\mathbf{e}_k^T}{1 + \frac{1}{n}\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k}. \tag{2.13}$$

Now write

$$\frac{1}{n}tr\mathbf{B}^{-1}(z) - E\frac{1}{n}tr\mathbf{B}^{-1}(z) = \frac{1}{n}\sum_{k=1}^{n}[E_k tr\mathbf{B}^{-1}(z) - E_{k-1}tr\mathbf{B}^{-1}(z)]$$

$$= \frac{1}{n}\sum_{k=1}^{n}(E_k - E_{k-1})(tr\mathbf{B}^{-1}(z) - tr\mathbf{B}_k^{-1}(z))$$

$$= \frac{1}{n}\sum_{k=1}^{n}(E_k - E_{k-1})\Big[\sum_{i=1}^{4} tr\Big(\mathbf{B}_k^{-1}(z)\mathbf{P}_y C_i \mathbf{P}_y \mathbf{B}^{-1}(z)\Big)\Big],$$

where the last step uses (2.10) and (2.12). Let $||\cdot||$ denote the spectral norm of matrices or the Euclidean norm of vectors. It is observed that

$$||\mathbf{B}^{-1}(z)|| \le \frac{1}{v}, \quad ||\mathbf{B}_k^{-1}(z)|| \le \frac{1}{v}, \quad ||\mathbf{P}_y|| \le 1, \quad \frac{1}{p_1}tr\mathbf{H}_k^{-1} \le \frac{1}{t}. \tag{2.14}$$

and since $\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k \ge 0$ we have

$$\frac{1}{1+\frac{1}{n}\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k} \le 1. \tag{2.15}$$

It follows that

$$\begin{aligned} |tr\mathbf{B}_k^{-1}(z)\mathbf{P}_y\mathbf{C}_1\mathbf{P}_y\mathbf{B}^{-1}(z)| &= \frac{1}{n^2}\Big|\frac{\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{X}_k\mathbf{P}_y\mathbf{B}^{-1}(z)\mathbf{B}_k^{-1}(z)\mathbf{P}_y\mathbf{X}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k}{1+\frac{1}{n}\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k}\Big| \\ \le \frac{1}{v^2n^2}||\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{X}_k||^2 &\le \frac{1}{v^2n}|\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k| + \frac{t}{v^2n}|\mathbf{x}_k^T\mathbf{H}_k^{-2}\mathbf{x}_k|, \end{aligned} \tag{2.16}$$

where the last inequality uses the facts that $||\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{X}_k||^2 = \mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{X}_k\mathbf{X}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k$ and $\mathbf{H}_k^{-1}\mathbf{X}_k\mathbf{X}_k^T\mathbf{H}_k^{-1} = n\mathbf{H}_k^{-1}(\frac{1}{n}\mathbf{X}_k\mathbf{X}_k^T + t\mathbf{I}_{p_1} - t\mathbf{I}_{p_1})\mathbf{H}_k^{-1} = n\mathbf{H}_k^{-1} - nt\mathbf{H}_k^{-2}$.

We then conclude from Lemma 2, (2.14)-(2.16) that

$$\begin{aligned} & E\Big|\frac{1}{n}\sum_{k=1}^{n}(E_k - E_{k-1})tr\mathbf{B}_k^{-1}(z)\mathbf{P}_y\mathbf{C}_1\mathbf{P}_y\mathbf{B}^{-1}(z)\Big|^4 \\ \le\ & \frac{M}{n^3}\sum_{k=1}^{n}E\Big|tr\mathbf{B}_k^{-1}(z)\mathbf{P}_y\mathbf{C}_1\mathbf{P}_y\mathbf{B}^{-1}(z)\Big|^4 \\ \le\ & \frac{M}{n^7}\sum_{k=1}^{n}E\Big|\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k\Big|^4 + \frac{M}{n^7}\sum_{k=1}^{n}E\Big|\mathbf{x}_k^T\mathbf{H}_k^{-2}\mathbf{x}_k\Big|^4 \\ =\ & O(\frac{1}{n^2}), \end{aligned}$$

where the last step uses the facts that via Lemma 3 and (2.9)

$$\frac{1}{n^4}E\Big|\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k\Big|^4 \le \frac{1}{n^4}ME\Big|\mathbf{x}_k^T\mathbf{H}_k^{-1}\mathbf{x}_k - tr\mathbf{H}_k^{-1}\Big|^4 + \frac{1}{n^4}ME|tr\mathbf{H}_k^{-1}|^4 \le M \tag{2.17}$$

and that

$$\frac{1}{n^4}E\Big|\mathbf{x}_k^T\mathbf{H}_k^{-2}\mathbf{x}_k\Big|^4 \le M. \tag{2.18}$$

Similarly, we can also obtain for $i = 2, 3, 4$,

$$E|\frac{1}{n}\sum_{k=1}^{n}(E_k - E_{k-1})tr\mathbf{B}^{-1}(z)\mathbf{P}_y\mathbf{C}_i\mathbf{P}_y\mathbf{B}_k^{-1}(z)|^4 \le \frac{M}{n^2}. \tag{2.19}$$

It follows from Borel-Cantelli's lemma that

$$\frac{1}{n}tr\mathbf{B}^{-1}(z) - E\frac{1}{n}tr\mathbf{B}^{-1}(z) \quad \text{a.s. } n \to \infty. \tag{2.20}$$

## 2.4 Step 4: From Gaussian distribution to general distributions

This section is to prove that

$$E[\frac{1}{n}tr\mathbf{B}^{-1}(z)] - E[\frac{1}{n}tr\mathbf{D}^{-1}(z)] \to 0 \;\; as \;\; n \to \infty, \tag{2.21}$$

where $\mathbf{D}^{-1}(z) = (\mathbf{P}_y\mathbf{P}_{tx}^g\mathbf{P}_y - z\mathbf{I})^{-1}$, $\mathbf{P}_{tx}^g = \frac{1}{n}\mathbf{G}^T(\frac{1}{n}\mathbf{G}\mathbf{G}^T + t\mathbf{I}_{p_1})^{-1}\mathbf{G}$ and $\mathbf{G} = (G_{ij})_{p_1\times n}$ consists of i.i.d. Gaussian random variables. We would point out that (2.4) follows immediately from (2.20), (2.21), tightness of $F^{\mathbf{B}}$ and the well-known inversion formula for Stieltjes transform[Theorem B.8 of Bai and Silverstein (2009)]. We use Lindeberg's method in Chatterjee (2006) to prove this result.

To facilitate statements, denote

$$X_{11}, \cdots, X_{1n}, X_{21}, \cdots, X_{p_1n} \; respectively \; by \; \hat{X}_1, \cdots, \hat{X}_n, \hat{X}_{n+1}, \cdots, \hat{X}_{p_1n}$$

and

$$G_{11}, \cdots, G_{1n}, G_{21}, \cdots, G_{p_1n} \; respectively \; by \; \hat{G}_1, \cdots, \hat{G}_n, \hat{G}_{n+1}, \cdots, \hat{G}_{p_1n}.$$

For each $j$, $0 \le j \le p_1n$, set

$$\mathbf{Z}_j = (\hat{X}_1, \cdots, \hat{X}_j, \hat{G}_{j+1}, \cdots, \hat{G}_{p_1n}) \; and \; \mathbf{Z}_j^0 = (\hat{X}_1, \cdots, \hat{X}_{j-1}, 0, \hat{G}_{j+1}, \cdots, \hat{G}_{p_1n}). \tag{2.22}$$

Note that $\mathbf{X}$ in $\mathbf{B}^{-1}(z)$ consists of the entries of $\mathbf{Z}_{p_1n}$. Hence we denote $\frac{1}{n}tr\mathbf{B}^{-1}(z)$ by $\frac{1}{n}tr(\mathbf{B}(\mathbf{Z}_{p_1n}) - zI)^{-1}$. Define the mapping $f$ from $R^{np_1}$ to $C$ as

$$f(\mathbf{Z}_{p_1n}) = \frac{1}{n}tr(\mathbf{B}(\mathbf{Z}_{p_1n}) - z\mathbf{I})^{-1}. \tag{2.23}$$

Furthermore we use the entries of $\mathbf{Z}_j$, $j = 0, 1, \cdots, p_1n - 1$, respectively, to replace $\hat{X}_1, \cdots, \hat{X}_{p_1n}$, the entries of $\mathbf{X}$ in $\mathbf{B}$, to constitute a series of new matrices. For these new matrices, we define $f(\mathbf{Z}_j)$, $j = 0, 1, \cdots, p_1n - 1$ as $f(\mathbf{Z}_{p_1n})$ is defined for the matrix $\mathbf{B}$. For example, $f(\mathbf{Z}_0) = \frac{1}{n}tr\mathbf{D}^{-1}(z)$. We then write

$$E[\frac{1}{n}tr\mathbf{B}^{-1}(z)] - E[\frac{1}{n}tr\mathbf{D}^{-1}(z)] = \sum_{j=1}^{p_1n}E\Big(f(\mathbf{Z}_j) - f(\mathbf{Z}_{j-1})\Big).$$

A third Taylor expansion yields

$$f(\mathbf{Z}_j) = f(\mathbf{Z}_j^0) + \hat{X}_j \partial_j f(\mathbf{Z}_j^0) + \frac{1}{2}\hat{X}_j^2 \partial_j^2 f(\mathbf{Z}_j^0) + \frac{1}{2}\hat{X}_j^3 \int_0^1 (1-\tau)^2 \partial_j^3 f(\mathbf{Z}_j^{(1)}(\tau))d\tau,$$

$$f(\mathbf{Z}_{j-1}) = f(\mathbf{Z}_j^0) + \hat{G}_j \partial_j f(\mathbf{Z}_j^0) + \frac{1}{2}\hat{G}_j^2 \partial_j^2 f(\mathbf{Z}_j^0) + \frac{1}{2}\hat{G}_j^3 \int_0^1 (1-\tau)^2 \partial_j^3 f(\mathbf{Z}_{j-1}^{(2)}(\tau))d\tau,$$

where $\partial_j^r f(\cdot)$, $r = 1, 2, 3$, stand for the $r$-fold derivative of the function $f$ in the $j$-th coordinate, and

$$\mathbf{Z}_j^{(1)}(\tilde{t}) = (\hat{X}_1, \cdots, \hat{X}_{j-1}, \tau \hat{X}_j, \hat{G}_{j+1}, \cdots, \hat{G}_{pn}),$$
$$\mathbf{Z}_{j-1}^{(2)}(\tilde{t}) = (\hat{X}_1, \cdots, \hat{X}_{j-1}, \tau \hat{G}_j, \hat{G}_{j+1}, \cdots, \hat{G}_{pn}).$$

Since $\hat{X}_j$ and $\hat{G}_j$ are both independent of $\mathbf{Z}_j^0$, $E[\hat{X}_j] = E[\hat{G}_j] = 0$ and $E[\hat{X}_j^2] = E[\hat{G}_j^2] = 1$, we obtain

$$\begin{aligned} & E[\frac{1}{n} tr \mathbf{B}^{-1}(z)] - E[\frac{1}{n} tr \mathbf{D}^{-1}(z)] \\ = \ & \frac{1}{2}\sum_{j=1}^{p_1 n} E\Big[\hat{X}_j^3 \int_0^1 (1-\tau)^2 \partial_j^3 f(\mathbf{Z}_j^{(1)}(\tau))d\tau - \hat{G}_j^3 \int_0^1 (1-\tau)^2 \partial_j^3 f(\mathbf{Z}_{j-1}^{(2)}(\tau))d\tau\Big]. \end{aligned}$$

Next we evaluate $\partial_j^3 f(\mathbf{Z}_{p_1 n}^{(1)}(\tau))$. Note that

$$\frac{\partial \mathbf{H}^{-1}}{\partial X_{ij}} = -\mathbf{H}^{-1}\frac{\partial \mathbf{H}}{\partial X_{ij}}\mathbf{H}^{-1}. \tag{2.24}$$

A simple but tedious calculation indicates that

$$\begin{aligned} \frac{\partial \mathbf{B}}{\partial X_{ij}} &= \frac{1}{n}\mathbf{P}_y \mathbf{e}_j \mathbf{e}_i^T \mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y + \frac{1}{n}\mathbf{P}_y \mathbf{X}^T \mathbf{H}^{-1}\mathbf{e}_i \mathbf{e}_j^T \mathbf{P}_y \\ &\quad -\frac{1}{n^2}\mathbf{P}_y \mathbf{X}^T \mathbf{H}^{-1}(\mathbf{e}_i \mathbf{e}_j^T \mathbf{X}^T + \mathbf{X}\mathbf{e}_j \mathbf{e}_i^T)\mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y, \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 \mathbf{B}}{\partial X_{ij}^2} &= \frac{2}{n}\mathbf{P}_y \mathbf{e}_j \mathbf{e}_i^T \mathbf{H}^{-1}\mathbf{e}_i \mathbf{e}_j^T \mathbf{P}_y - \frac{2}{n^2}\mathbf{P}_y \mathbf{e}_j \mathbf{e}_i^T \mathbf{H}^{-1}(\mathbf{e}_i \mathbf{e}_j^T \mathbf{X}^T + \mathbf{X}\mathbf{e}_j \mathbf{e}_i^T)\mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y \\ &\quad -\frac{2}{n^2}\mathbf{P}_y \mathbf{X}^T \mathbf{H}^{-1}(\mathbf{e}_i \mathbf{e}_j^T \mathbf{X}^T + \mathbf{X}\mathbf{e}_j \mathbf{e}_i^T)\mathbf{H}^{-1}\mathbf{e}_i \mathbf{e}_j^T \mathbf{P}_y - \frac{2}{n^2}\mathbf{P}_y \mathbf{X}^T \mathbf{H}^{-1}\mathbf{e}_i \mathbf{e}_i^T \mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y \\ &\quad +\frac{2}{n^3}\mathbf{P}_y \mathbf{X}^T [\mathbf{H}^{-1}(\mathbf{e}_i \mathbf{e}_j^T \mathbf{X}^T + \mathbf{X}\mathbf{e}_j \mathbf{e}_i^T)]^2 \mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y, \end{aligned}$$

$$\begin{aligned}
\frac{\partial^3 \mathbf{B}}{\partial X_{ij}^3} &= -\frac{6}{n^2}\mathbf{P}_y\mathbf{e}_j\mathbf{e}_i^T\mathbf{H}^{-1}(\mathbf{e}_i\mathbf{e}_j^T\mathbf{X}^T+\mathbf{X}\mathbf{e}_j\mathbf{e}_i^T)\mathbf{H}^{-1}\mathbf{e}_i\mathbf{e}_j^T\mathbf{P}_y - \frac{6}{n^2}\mathbf{P}_y\mathbf{e}_j\mathbf{e}_i^T\mathbf{H}^{-1}\mathbf{e}_i\mathbf{e}_i^T\mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y \\
&+\frac{6}{n^3}\mathbf{P}_y\mathbf{e}_j\mathbf{e}_i^T[\mathbf{H}^{-1}(\mathbf{e}_i\mathbf{e}_j^T\mathbf{X}^T+\mathbf{X}\mathbf{e}_j\mathbf{e}_i^T)]^2\mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y - \frac{6}{n^2}\mathbf{P}_y\mathbf{X}^T\mathbf{H}^{-1}\mathbf{e}_i\mathbf{e}_i^T\mathbf{H}^{-1}\mathbf{e}_i\mathbf{e}_j^T\mathbf{P}_y \\
&+\frac{6}{n^3}\mathbf{P}_y\mathbf{X}^T[\mathbf{H}^{-1}(\mathbf{e}_i\mathbf{e}_j^T\mathbf{X}^T+\mathbf{X}\mathbf{e}_j\mathbf{e}_i^T)]^2\mathbf{H}^{-1}\mathbf{e}_i\mathbf{e}_j^T\mathbf{P}_y \\
&-\frac{6}{n^4}\mathbf{P}_y\mathbf{X}^T[\mathbf{H}^{-1}(\mathbf{e}_i\mathbf{e}_j^T\mathbf{X}^T+\mathbf{X}\mathbf{e}_j\mathbf{e}_i^T)]^3\mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y \\
&+\frac{6}{n^3}\mathbf{P}_y\mathbf{X}^T\mathbf{H}^{-1}(\mathbf{e}_i\mathbf{e}_j^T\mathbf{X}^T+\mathbf{X}\mathbf{e}_j\mathbf{e}_i^T)\mathbf{H}^{-1}\mathbf{e}_i\mathbf{e}_i^T\mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y \\
&+\frac{6}{n^3}\mathbf{P}_y\mathbf{X}^T\mathbf{H}^{-1}\mathbf{e}_i\mathbf{e}_i^T\mathbf{H}^{-1}(\mathbf{e}_i\mathbf{e}_j^T\mathbf{X}^T+\mathbf{X}\mathbf{e}_j\mathbf{e}_i^T)\mathbf{H}^{-1}\mathbf{X}\mathbf{P}_y.
\end{aligned}$$

Also, by the formula

$$\frac{1}{n}\frac{\partial tr\mathbf{B}^{-1}(z)}{\partial X_{ij}} = -\frac{1}{n}tr(\frac{\partial \mathbf{B}}{\partial X_{ij}}\mathbf{B}^{-2}(z)),$$

it is easily seen that

$$\begin{aligned}
\frac{1}{n}\frac{\partial^3 tr\mathbf{B}^{-1}(z)}{\partial X_{ij}^3} &= -\frac{6}{n}tr(\frac{\partial \mathbf{B}}{\partial X_{ij}}\mathbf{B}^{-1}(z)\frac{\partial \mathbf{B}}{\partial X_{ij}}\mathbf{B}^{-1}(z)\frac{\partial \mathbf{B}}{\partial X_{ij}}\mathbf{B}^{-2}(z)) \\
&-\frac{1}{n}tr(\frac{\partial^3 \mathbf{B}}{\partial X_{ij}^3}\mathbf{B}^{-2}(z)) + \frac{3}{n}tr(\frac{\partial^2 \mathbf{B}}{\partial X_{ij}^2}\mathbf{B}^{-2}(z)\frac{\partial \mathbf{B}}{\partial X_{ij}}\mathbf{B}^{-1}(z)) \\
&+\frac{3}{n}tr(\frac{\partial^2 \mathbf{B}}{\partial X_{ij}^2}\mathbf{B}^{-1}(z)\frac{\partial \mathbf{B}}{\partial X_{ij}}\mathbf{B}^{-2}(z)).
\end{aligned}$$

There are lots of terms in the expansion of $\frac{1}{n}\frac{\partial^3 tr\mathbf{B}^{-1}(z)}{\partial X_{ij}^3}$ and therefore we do not enumerate all the terms here. By using the formula that, for any matrices $\mathbf{A}$, $\mathbf{B}$ and column vectors $\mathbf{e}_j$ and $\mathbf{e}_k$,

$$tr(\mathbf{A}\mathbf{e}_j\mathbf{e}_k^T\mathbf{B}) = \mathbf{e}_k^T\mathbf{B}\mathbf{A}\mathbf{e}_j, \tag{2.25}$$

all the terms of $\frac{1}{n}\frac{\partial^3 tr\mathbf{B}^{-1}(z)}{\partial X_{ij}^3}$ can be dominated by a common expression. That is

$$\begin{aligned}
||\frac{1}{n}\frac{\partial^3 tr\mathbf{B}^{-1}(z)}{\partial X_{ij}^3}|| &\le \frac{M}{n^3}||\mathbf{H}^{-1}||\cdot||\mathbf{X}^T\mathbf{H}^{-1}|| + \frac{M}{n^4}||\mathbf{X}^T\mathbf{H}^{-1}||^3 \\
&+\frac{M}{n^4}||\mathbf{H}^{-1}||\cdot||\mathbf{X}^T\mathbf{H}^{-1}||^2 \\
&+\frac{M}{n^4}||\mathbf{H}^{-1}||\cdot||\mathbf{X}^T\mathbf{H}^{-1}||\cdot||\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}|| \\
&+\frac{M}{n^5}||\mathbf{H}^{-1}||\cdot||\mathbf{X}^T\mathbf{H}^{-1}||\cdot||\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}||^2 \\
&+\frac{M}{n^5}||\mathbf{X}^T\mathbf{H}^{-1}||^3\cdot||\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}|| \\
&+\frac{M}{n^6}||\mathbf{X}^T\mathbf{H}^{-1}||^3\cdot||\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}||^2 \\
&+\frac{M}{n^7}||\mathbf{X}^T\mathbf{H}^{-1}||^3\cdot||\mathbf{X}^T\mathbf{H}\mathbf{X}||^3.
\end{aligned} \tag{2.26}$$

Obviously
$$||\mathbf{H}^{-1}|| \le \frac{1}{t}. \tag{2.27}$$
It is observed that
$$\begin{aligned} ||\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}||^2 &= \lambda_{\max}(\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}) = \lambda_{\max}(\mathbf{H}^{-1}\mathbf{X}\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}\mathbf{X}^T) \\ &\le n^2[1+2t||\mathbf{H}^{-1}|| + t^2||\mathbf{H}^{-2}||] \le Mn^2, \end{aligned} \tag{2.28}$$
where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of the corresponding matrix; and the first inequality above utilizes the fact that $\mathbf{H}^{-1}\mathbf{X}\mathbf{X}^T = n\mathbf{H}^{-1}(\frac{1}{n}\mathbf{X}\mathbf{X}^T + t\mathbf{I}_{p_1} - t\mathbf{I}_{p_1}) = n\mathbf{I}_{p_1} - nt\mathbf{H}^{-1}$.

Similarly we can obtain
$$||\mathbf{X}^T\mathbf{H}^{-1}|| \le M\sqrt{n}. \tag{2.29}$$
We conclude from (2.26)-(2.29) that
$$||\frac{1}{n}\frac{\partial^3 tr\mathbf{B}^{-1}(z)}{\partial X_{ij}^3}|| \le \frac{M}{n^{5/2}}. \tag{2.30}$$
This implies that
$$E|X_{ij}^3 \cdot \frac{1}{n}\frac{\partial^3 tr\mathbf{B}^{-1}(z)}{\partial X_{ij}^3}| \le \frac{M}{n^{5/2}}E[X_{ij}^3] \le \frac{M\varepsilon_n}{n^2}. \tag{2.31}$$
Since all $X_{ij}$ and $W_{ij}$ play a similar role in their corresponding matrices, the above argument works for all matrices. Hence we obtain
$$\begin{aligned} & |E[\frac{1}{n}tr\mathbf{B}^{-1}(z)] - E[\frac{1}{n}tr\mathbf{D}^{-1}(z)]| \\ \le\ & M\sum_{j=1}^{p_1 n}[\int_0^1 (1-\tau)^2 E|\hat{X}_j^3 \partial_j^3 f(\mathbf{Z}_j^{(1)}(\tau))|d\tau + \int_0^1 (1-\tau)^2 E|\hat{G}_j^3 \partial_j^3 f(\mathbf{Z}_{j-1}^{(2)}(\tau))|d\tau] \\ \le\ & M\varepsilon_n. \end{aligned}$$
This ensures that
$$E[\frac{1}{n}tr\mathbf{B}^{-1}(z)] - E[\frac{1}{n}tr\mathbf{D}^{-1}(z)] \to 0 \quad as \quad n \to \infty.$$
Therefore the proof of Theorem 1 is completed.

# 3 Conclusion

Canonical correlation coefficients play an important role in the analysis of correlations between random vectors[Anderson (1984)]. Nowadays, investigations of large dimensional random vectors attract a substantial research works, e.g. Fan and Lv (2010). As future works, we plan to develop central limit theorems for the empirical distribution of canonical correlation coefficients and make statistical applications of the developed asymptotic theorems for large dimensional random vectors.

# 4 Appendix

**Lemma 1** (Burkholder (1973)). *Let $\{X_k, 1 \le k \le n\}$ be a complex martingale difference sequence with respect to the increasing $\sigma$-field $\{\mathcal{F}_k\}$. Then, for $p \ge 2$,*

$$E|\sum_{k=1}^{n} X_k|^p \le K_p(E(\sum_{k=1}^{n} E(|X_k|^2|\mathcal{F}_{k-1}))^{p/2} + E\sum_{k=1}^{n} |X_k|^p).$$

**Lemma 2** (Burkholder (1973)). *With $\{X_k, 1 \le k \le n\}$ as above, we have, for $p > 1$,*

$$E|\sum_{k=1}^{n} X_k|^p \le K_p E(\sum_{k=1}^{n} |X_k|^2)^{p/2}.$$

**Lemma 3** (Lemma B.26 of Bai and Silverstein (2009)). *For $\mathbf{X} = (X_1, \cdots, X_n)^T$ i.i.d standardized entries, $\mathbf{C}$ $n \times n$ matrix, we have, for any $p \ge 2$,*

$$E|\mathbf{X}^*\mathbf{C}\mathbf{X} - tr\mathbf{C}|^p \le K_p((E|X_1|^4 tr\mathbf{C}\mathbf{C}^*)^{p/2} + E|X_1|^{2p} tr(\mathbf{C}\mathbf{C}^*)^{p/2}).$$

**Lemma 4** (Theorem A.43 of Bai and Silverstein (2009)). *Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ symmetric matrices. Then*

$$||F^{\mathbf{A}} - F^{\mathbf{B}}|| \le \frac{1}{n} rank(\mathbf{A} - \mathbf{B}),$$

*where $||f|| = sup_x|f(x)|$.*

**Lemma 5** (Hoeffding (1963)). *Let $Y_1, Y_2, \ldots$ be i.i.d random variables, $P(Y_1 = 1) = q = 1 - P(Y_1 = 0)$. Then*

$$P(|Y_1 + \cdots + Y_n - nq| \ge n\varepsilon) \le 2e^{-\frac{n^2\varepsilon^2}{2nq + n\varepsilon}}$$

*for all $\varepsilon > 0$, $n = 1, 2, \ldots$.*

**Lemma 6** (Corollary A.41 of Bai and Silverstein (2009)). *Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ symmetric matrices with their respective ESDs of $F^{\boldsymbol{A}}$ and $F^{\boldsymbol{B}}$. Then,*

$$L^3(F^{\boldsymbol{A}}, F^{\boldsymbol{B}}) \le \frac{1}{n} tr(\boldsymbol{A} - \boldsymbol{B})^2.$$

# References


T.W. Anderson, An introduction to multivariate statistical analysis, *New York: Wiley*, **2th edtion** (1984).

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

Z.D. Bai, J.W. Silverstein, Spectral analysis of large dimensional random matrices, *Springer*, **2nd edition** (2009).

W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.*, **58(301)** (1963) 13-30.

D.L. Burkholder, Distribution function inequalities for martingales, *Ann. Probab.*, **1(1)** (1973) 19-42.

S. Chatterjee, A generalization of the Lindeberg principle, *Ann. Probab.*, **34(6)** (2006) 2061-2076.

J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, *Statist. Sinica*, **20(1)** (2010) 101-148.

I.M. Johnstone, Multivariate analysis and Jacobi ensembles: largest eigenvalue, tracy-widom limits and rates of convergence, *Ann. Statist.*, **36(6)** (2008) 2638-2716.

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

K.V. Mardia, J.T. Kent, and J.M. Bibby, Multivariate Analysis, *Academic Press* (1979), London. MR0560319.

J.W. Lindeberg, Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung, *Math. Z.*, **15** (1922) 211–225.

V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, *Math. USSR-Sb.*, **1(4)** (1967) 457-483.

G.M. Pan, Strong convergence of the empirical distribution of eigenvalues of sample covariance matrices with a perturbation matrix, *J. Multivariate Anal.*, **101(6)** (2010) 1330-1338.

K.W. Wachter, The limiting empirical measure of multiple discriminant ratios, *Ann. Statist.*, **8(5)** (1980) 937-957.

Y. R. Yang and G. M. Pan, Independence Test For High Dimensional Data Based On Regularized Canonical Correlation Coefficients. Preprint (2012)).

Y.Q. Yin, Limiting spectral distribution for a class of random matrices, *J. Multivariate Anal.*, **20(1)** (1986) 50-68.