Statistica Sinica **9**(1999), 611-677

METHODOLOGIES IN SPECTRAL ANALYSIS OF LARGE DIMENSIONAL RANDOM MATRICES, A REVIEW

Z. D. Bai

National University of Singapore

Abstract: In this paper, we give a brief review of the theory of spectral analysis of large dimensional random matrices. Most of the existing work in the literature has been stated for real matrices but the corresponding results for the complex case are also of interest, especially for researchers in Electrical and Electronic Engineering. Thus, we convert almost all results to the complex case, whenever possible. Only the latest results, including some new ones, are stated as theorems here. The main purpose of the paper is to show how important methodologies, or mathematical tools, have helped to develop the theory. Some unsolved problems are also stated.

Key words and phrases: Circular law, complex random matrix, largest and smallest eigenvalues of a random matrix, noncentral Hermitian matrix, spectral analysis of large dimensional random matrices, spectral radius.

1. Introduction

The necessity of studying the spectra of LDRM (Large Dimensional Random Matrices), especially the Wigner matrices, arose in nuclear physics during the 1950's. In quantum mechanics, the energy levels of quantums are not directly observable, but can be characterized by the eigenvalues of a matrix of observations. However, the ESD (Empirical Spectral Distribution, the empirical distribution of the eigenvalues) of a random matrix has a very complicated form when the order of the matrix is high. Many conjectures, e.g., the famous circular law, were made through numerical computation.

Since then, research on the LSA (Limiting Spectral Analysis) of LDRM has attracted considerable interest among mathematicians, probabilitists and statisticians. One pioneering work is Wigner's semicircular law for a Gaussian (or Wigner) matrix (Wigner (1955, 1958)). He proved that the expected ESD of a large dimensional Wigner matrix tends to the so-called semicircular law. This work was generalized by Arnold (1967, 1971) and Grenander (1963) in various aspects. Bai and Yin (1988a) proved that the ESD of a suitably normalized sample covariance matrix tends to the semicircular law when the dimension is relatively smaller than the sample size. Following the work by Marčenko and Pastur (1967) and Pastur (1972, 1973), the LSA of large dimensional sample covariance matrices was developed by many researchers, including Bai, Yin and

Krishnaiah (1986), Grenander and Silverstein (1977), Jonsson (1982), Wachter (1978), Yin (1986) and Yin and Krishnaiah (1983). Also, Bai, Yin and Krishnaiah (1986, 1987), Silverstein (1985a), Wachter (1980), Yin (1986) and Yin and Krishnaiah (1983) investigated the LSD (Limiting Spectral Distribution) of the multivariate F-matrix, or more generally, of products of random matrices.

Two important problems arose after the LSD was found: bounds on extreme eigenvalues and the convergence rate of the ESD. The literature on the first problem is extensive. The first success was due to Geman (1980), who proved that the largest eigenvalue of a sample covariance matrix converges almost surely to a limit under a growth condition on all the moments of the underlying distribution. Yin, Bai and Krishnaiah (1988) proved the same result under the existence of the fourth moment, and Bai, Silverstein and Yin (1988) proved that the existence of the fourth moment is also necessary for the existence of the limit. Bai and Yin (1988b) found necessary and sufficient conditions for the almost sure convergence of the largest eigenvalue of a Wigner matrix. Bai and Yin (1993), Silverstein (1985b) and Yin, Bai and Krishnaiah (1983) considered the almost sure limit of the smallest eigenvalue of a sample covariance matrix. Some related work can be found in Geman (1986) and Bai and Yin (1986).

The second problem, the convergence rate of the ESD's of LDRM, is of practical interest, but had been open for decades. The first success was made in Bai (1993a,b), in which convergence rates for the expected ESD of a large Wigner matrix and sample covariance matrix were established. Further extensions of this work can be found in Bai, Miao and Tsay (1996a,b, 1997).

The most perplexing problem is the so-called circular law that conjectures that the ESD of a non-symmetric random matrix, after suitable normalization, tends to the uniform distribution over the unit disc in the complex plane. The difficulty lies in the fact that two most important tools for symmetric matrices do not apply to non-symmetric matrices. Furthermore, certain truncation and centralization techniques cannot be used. The first known result, a partial solution for matrices whose entries are i.i.d. standard complex normal (whose real and imaginary parts are i.i.d. real normal with mean zero and variance $1/2$), was given in Mehta (1991) and an unpublished result of Silverstein reported in Hwang (1986). They used the explicit expression of the joint density of the complex eigenvalues of a matrix with independent standard complex Gaussian entries, found by Ginibre (1965). The first attempt to prove this conjecture under some general conditions was made in Girko (1984a,b). However, his proofs have puzzled many who attempted to understand the arguments. Recently, Edelman (1997) found the joint distribution of the eigenvalues of a matrix whose entries are real normal $N(0,1)$ and proved that the expected ESD of a matrix of i.i.d. real Gaussian entries tends to the circular law. Under the existence of the $(4+\varepsilon)$th moment and some smoothness conditions, Bai (1997) proved the strong version of the circular law.

In this paper, we give a brief review of the main results in this area and some of their applications. We convert most known results for real matrices to complex matrices. Since some of the extensions are non-trivial, we give brief outlines of their proofs. The review will be given in accordance with methodologies by which the results were obtained, rather than in a chronological order. The organization of the paper is as follows. In Section 2, we give results obtained by the moment approach. In Section 3, the Stieltjes transform technique is introduced. Recent achievements on the circular law are presented in Section 4. Some applications are mentioned in Section 5, and some open problems or conjectures are presented in Section 6.

2. Moment Approach

Throughout this section, we consider only Hermitian matrices which include real symmetric matrices as a special case. Let $\mathbf{A}$ be a $p \times p$ Hermitian matrix and denote its eigenvalues by $\lambda_1 \leq \cdots \leq \lambda_p$. The ESD of $\mathbf{A}$ is defined by $F^{\mathbf{A}}(x) = p^{-1}\#\{k \leq p,\ \lambda_k \leq x\}$, where $\#\{.\}$ denotes the number of elements in the set indicated. A simple fact is that the hth moment of $F^{\mathbf{A}}$ can be written as

$$\beta_h(\mathbf{A}) = \int x^h F^{\mathbf{A}}(dx) = p^{-1}\mathrm{tr}(\mathbf{A}^h). \tag{2.1}$$

This formula plays a fundamental role in the theory of LDRM. Most of the results were obtained by estimating the mean, variance or higher moments of $p^{-1}\mathrm{tr}(\mathbf{A}^h)$.

2.1. Limiting spectral distributions

To show that $F^{\mathbf{A}}$ tends to a limit, say F, usually the Moment Convergence Theorem (MCT) is employed, i.e., verifying

$$\beta_h(\mathbf{A}) \to \beta_h = \int x^h F(dx)$$

in some sense (e.g., almost surely (a.s.) or in probability) and the Carleman's condition $\sum_{h=1}^{\infty} \beta_{2h}^{-1/(2h)} = \infty$. Note that Carleman's condition is slightly weaker than requiring that the characteristic function of the LSD be analytic near 0. For most cases, the LSD has bounded support and its characteristic function is analytic. Thus, the MCT can be applied to show the existence of the LSD of a sequence of random matrices.

2.1.1. Wigner matrix

In this subsection, we first consider the famous semicircular law. A Wigner matrix $\mathbf{W}$ of order n is defined as an $n \times n$ Hermitian matrix whose entries above the diagonal are i.i.d. complex random variables with variance σ^2, and whose diagonal elements are i.i.d. real random variables (without any moment requirement). We have the following theorem.

Theorem 2.1. *Under the conditions described above, as $n \to \infty$, with probability 1, the ESD $F^{(n^{-1/2}\mathbf{W})}$ tends to the semicircular law with scale-parameter σ, whose density is given by*

$$p_\sigma(x) = \begin{cases} \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - x^2}, & \text{if } |x| \le 2\sigma, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

The proof of the theorem consists of centralization, truncation and convergence of moments. First, we present two lemmas. Throughout the paper, $\|f\| = \sup_x |f(x)|$.

Lemma 2.2 (Rank Inequality). *Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ Hermitian matrices. Then*

$$\|F^{\mathbf{A}} - F^{\mathbf{B}}\| \le \frac{1}{n}\text{rank}(\mathbf{A} - \mathbf{B}). \tag{2.3}$$

Suppose that $\text{rank}(\mathbf{A} - \mathbf{B}) = k$. To prove (2.3), we may assume that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where the order of $\mathbf{A}_{22}$ is $(n-k) \times (n-k)$, since (2.3) is invariant under unitary similarity. Denote the eigenvalues of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A}_{22}$ by $\lambda_1 \le \cdots \le \lambda_n$, $\eta_1 \le \cdots \le \eta_n$ and $\tilde\lambda_1 \le \cdots \le \tilde\lambda_{(n-k)}$, respectively. By the interlacing inequality $\max(\lambda_j, \eta_j) \le \tilde\lambda_j \le \min(\lambda_{(j+k)}, \eta_{(j+k)})$ (see Rao (1976, p.64), referred to as *Poincare Separation Theorem*), we conclude that for any $x \in (\tilde\lambda_{(j-1)}, \tilde\lambda_j)$, $\frac{j-1}{n} \le F^{\mathbf{A}}(x),\ F^{\mathbf{B}}(x) < \frac{j+k}{n}$, which implies (2.3).

Lemma 2.3 (Difference Inequality). *Let $\mathbf{A}$ and $\mathbf{B}$ be two $n \times n$ complex normal matrices with complex eigenvalues $\lambda_1, \ldots, \lambda_n$ and $\eta_1, \ldots, \eta_n$, respectively. Then*

$$\min_\pi \sum_{k=1}^n |\lambda_k - \eta_{\pi_k}|^2 \le \text{tr}(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^*, \tag{2.4}$$

where π is a permutation of the set $\{1, \ldots, n\}$. Consequently, if both $\mathbf{A}$ and $\mathbf{B}$ are Hermitian, then

$$L^3(F^{\mathbf{A}}, F^{\mathbf{B}}) \le \frac{1}{n}\text{tr}(\mathbf{A} - \mathbf{B})^2 \tag{2.5}$$

where $L(F, G)$ denotes the Levy distance between distribution functions F and G. This lemma is an improvement to Lemma 2.3 *of Bai* (1993b), *where the power of the Levy distance is* 4.

To prove (2.4), we may assume, without loss of generality, that $\mathbf{A} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\mathbf{B} = \mathbf{U}^*\mathrm{diag}(\eta_1, \ldots, \eta_n)\mathbf{U}$, where $\mathbf{U} = (u_{jk})$ is a unitary matrix. Then we have $\mathrm{tr}(\mathbf{AA}^*) = \sum_{k=1}^n |\lambda_k^2|$, $\mathrm{tr}(\mathbf{BB}^*) = \sum_{k=1}^n |\eta_k^2|$ and for some permutation π, $\mathrm{Re}(\mathrm{tr}(\mathbf{AB}^*)) = \sum_{jk} \mathrm{Re}(\lambda_j \bar\eta_k)|u_{jk}^2| \le \sum_{k=1}^n \mathrm{Re}(\lambda_k \bar\eta_{\pi_k})$. The last inequality holds since the maximum of $\sum_{jk} \mathrm{Re}(\lambda_j\bar\eta_k)a_{jk}$, subject to the constraints $a_{jk} \ge 0$ and $\sum_j a_{jk} = \sum_k a_{jk} = 1$, is equal to $\sum_{k=1}^n \mathrm{Re}(\lambda_k\bar\eta_{\pi_k})$ for some permutation π. Then (2.4) follows.

If $\mathbf{A}$ and $\mathbf{B}$ are both Hermitian, their eigenvalues λ_k's and η_k's are real and hence can be arranged in increasing order. In this case, the solution of the minimization on the left hand side of (2.4) is $\pi_k = k$. One can show that

$$L^3(F^{\mathbf{A}}, F^{\mathbf{B}}) \le \frac{1}{n}\sum_{k=1}^n |\lambda_k - \eta_k|^2,$$

which, together with (2.4), implies (2.5). The proof of the lemma is then complete.

Now, we outline a proof of Theorem 2.1. Note that $\mathrm{E}(\mathrm{Re}(w_{jk})) = \mathrm{E}(\mathrm{Re}(w_{kj}))$. Applying Lemma 2.2, we conclude that the LSD's of $n^{-1/2}\mathbf{W}$ and $n^{-1/2}(\mathbf{W} - \mathrm{E}(\mathrm{Re}(w_{12}))\mathbf{11}')$ exist and are the same if either one of them exists. Thus, we can remove the real parts of the expectation of off-diagonal entries (i.e., to replace them by 0). Further, note that the eigenvalues of $i\mathrm{E}(\mathrm{Im}(n^{-1/2}\mathbf{W}))$ are given by $n^{-1/2}\mathrm{E}(\mathrm{Im}(w_{12}))\cot(\pi(2k-1)/2n)$, $k = 1, \ldots, n$. Thus, applying Lemma 2.2, we can remove the eigenvalues of $i\mathrm{E}(\mathrm{Im}(n^{-1/2}\mathbf{W}))$ whose absolute values are greater than $n^{-1/4}$ (the number of such eigenvalues is less than $4n^{3/4}|E\mathrm{Im}(w_{12})|$), and by Lemma 2.3, we can also remove the remaining eigenvalues. Therefore, we may assume that the means of the off-diagonal entries of $n^{-1/2}\mathbf{W}$ are zero. Now we remove the diagonal elements of $\mathbf{W}$ by employing Hoeffding's (1963) inequality, which states that for any $\varepsilon > 0$,

$$\mathrm{P}(|\xi_n - \eta| \ge \varepsilon) \le 2\exp(-\varepsilon^2/(2\eta + \varepsilon)),$$

where $\eta = \mathrm{E}(\xi_n)$ and ξ_n is the sum of n independent random variables taking values 0 or 1 only.

By this inequality and the fact that $\mathrm{P}(|w_{11}| \ge \varepsilon\sqrt{n}) = o(1)$, we have

$$\mathrm{P}(\sum_{k=1}^n I_{\{|n^{-1/2}w_{kk}|\ge\varepsilon\}} \ge \varepsilon n) \le \mathrm{P}(\sum_{k=1}^n \Big[I_{\{|n^{-1/2}w_{kk}|\ge\varepsilon\}} - \mathrm{P}(|n^{-1/2}w_{11}| \ge \varepsilon)\Big] \ge \frac{1}{2}\varepsilon n)$$
$$\le 2\exp\{-n^2/(2n\mathrm{P}(|w_{11}| \ge \varepsilon\sqrt{n}) + \varepsilon n)\} \le 2e^{-bn}$$

for some $b > 0$ and all large n. Applying Lemma 2.2, one can remove the diagonal entries greater than ε and, by Lemma 2.3, can also remove those smaller than ε without altering the limiting distribution of $n^{-1/2}\mathbf{W}$.

Now we use the truncation and centralization techniques. Let $\widetilde{\mathbf{W}}$ be the matrix with zero diagonal entries and off-diagonal entries $w_{jk}I_{(|w_{jk}|\le C)} - \mathrm{E}(w_{jk} I_{(|w_{jk}|\le C)})$. By Lemma 2.3 and the Law of Large Numbers, with probability 1,

$$\begin{aligned} &L^4(F^{(n^{-1/2}\mathbf{W})}, F^{(n^{-1/2}\widetilde{\mathbf{W}})}) \\ &\le n^{-2}\sum_{j,k}|w_{jk}I_{(|w_{jk}|>C)} - \mathrm{E}(w_{jk}I_{(|w_{jk}|>C)})|^2 \\ &\to \mathrm{E}|w_{12}I_{(|w_{12}|>C)} - \mathrm{E}(w_{12}I_{(|w_{12}|>C)})|^2 \le \mathrm{E}|w_{12}I_{(|w_{12}|>C)}|^2. \end{aligned} \tag{2.6}$$

Note that $\mathrm{E}|w_{12}I_{(|w_{12}|>C)}|^2$ can be arbitrarily small if C is large enough. Thus, in the proof of Theorem 2.1, we may assume that the entries of $\mathbf{W}$ are bounded by C.

Next we establish the convergence of moments of the ESD of $n^{-1/2}\mathbf{W}$, see (2.1). For given integers $j_1,\ldots,j_h(\le n)$, construct a W-graph G by plotting $j_1,\ldots,j_h$ on a straight line as vertices and drawing h edges from j_r to j_{r+1} (with $j_{h+1} = j_1$), $r = 1,\ldots,h$. An example of a W-graph is given in Figure 1, in which there are 10 vertices (i_1-i_{10}), 4 non-coincident vertices (v_1-v_4), 9 edges, 4 non-coincident edges and 1 single edge (v_4, v_3). For this graph, we say the edge (v_1, v_2) has multiplicity 2 and the edges (v_2, v_3) and (v_2, v_4) have multiplicities 3. An edge (u, v) corresponds to the variable $w_{u,v}$ and a W-graph corresponds to the product of variables corresponding to the edges making up this W-graph.

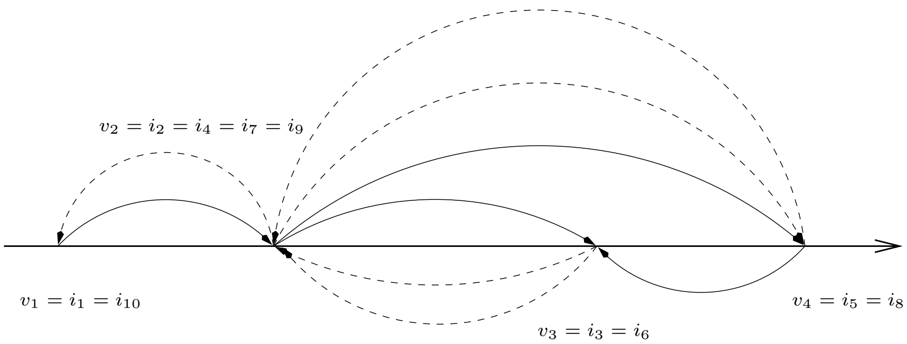

Figure 1. Types of edges in a W-graph.

Note that

$$\frac{1}{n}\mathrm{tr}((n^{-1/2}\mathbf{W})^h) = n^{-1-h/2}\sum_{j_1,\ldots,j_h} w_{j_1j_2}w_{j_2,j_3}\ldots w_{j_hj_1} := n^{-1-h/2}\sum_G w_G.$$

Then Theorem 2.1 follows by showing that

$$\frac{1}{n}\mathrm{E}(\mathrm{tr}((n^{-1/2}\mathbf{W})^h)) = n^{-1-h/2}\sum_G \mathrm{E}(w_G) = \begin{cases} \frac{(2s)!\sigma^h}{s!(s+1)!} + O(n^{-1}), & \text{if } h = 2s \\ O(n^{-1/2}), & \text{if } h = 2s+1 \end{cases} \tag{2.7}$$

and

$$\mathrm{Var}(\frac{1}{n}\mathrm{tr}((n^{-1/2}\mathbf{W})^h)) = n^{-2-h}\sum_{G_1,G_2}[\mathrm{E}(w_{G_1}w_{G_2}) - \mathrm{E}(w_{G_1})\mathrm{E}(w_{G_2})] = O(n^{-2}), \tag{2.8}$$

through the following arguments.

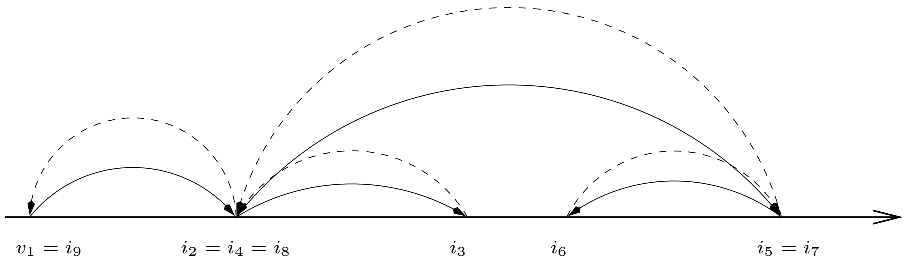

Figure 2. A W-graph with 8 edges, 4 non-coincident edges and 5 vertices.

To prove (2.7), we note that if there is a single edge in the W-graph, the corresponding expectation is zero. When $h = 2s+1$, there are at most s non-coincident edges and hence at most $s+1$ non-coincident vertices. This shows that there are at most n^{s+1} graphs (or non-zero terms in the expansion). Then the second conclusion in (2.7) follows since the denominator is $n^{s+3/2}$ and the absolute value of the expectation of each term is not larger than C^h. When $h = 2s$, classify the graphs into two types. The first type, consisting of all graphs which have at most s non-coincident vertices, gives the estimation of the remainder term $O(n^{-1})$. The second type consists of all graphs which have exactly $s+1$ non-coincident vertices and s non-coincident edges. There are no cycles of non-coincident edges in such graphs and each edge (u,v) must coincide with and only with the edge (v,u) which corresponds to $\mathrm{E}|w_{uv}|^2 = \sigma^2$. Thus, each term corresponding to a second type W-graph is σ^h. To complete the proof of the first conclusion of (2.7), we only need to count the number of second type W-graphs. We say that two W-graphs are isomorphic if one can be coverted to the other by a permutation of $\{1,\ldots,n\}$ on the straight line. We first compute the number of isomorphic classes. If an edge (u,v) is single in the subgraph $[(i_1,i_2),\ldots,(i_t,i_{t+1})]$, we say that this edge is single up to the edge (i_t,i_{t+1}) or the vertex i_{t+1}. In a second type W-graph, there are s edges which are single up to themselves, and the other s edges coincide with existing edges when they are drawn. Define a CS (Characteristic Sequence) for a W-graph by $u_i = 1$ (or -1) if the ith edge is single up to its end vertex (or coinciding with an existing edge, respectively). For example, for the graph in Figure 2, its CS is $1, 1, -1, 1, 1, -1, -1, -1$. The sequence $u_1, \ldots, u_{2s}$ should satisfy $\sum_{i=1}^{j} u_i \geq 0$, for all $j = 1, \ldots, 2s$.

The number of all arrangements of the ± 1's is $\frac{(2s)!}{s!s!}$. By the reflection principle (see Figure 3), the number of arrangements such that at least one of the requirements $\sum_{i=1}^{j} u_i \geq 0$ is violated is $\frac{(2s)!}{(s-1)!(s+1)!}$ (see the broken curve which reaches the line $y = -1$; reflecting the rear part of the curve across the axis $y = -1$ results in the dotted curve which ends at $y = -2$ and consists of $s - 1$ ones and $s + 1$ minus ones). It follows that the number of isomorphic classes is $\frac{(2s)!}{s!s!} - \frac{(2s)!}{(s-1)!(s+1)!} = \frac{(2s)!}{s!(s+1)!}$. The number of graphs in each isomorphic class is $n(n-1) \cdots (n-s) = n^{1+s}(1 + O(n^{-1}))$. Then the first conclusion in (2.7) follows. The proof of (2.8) follows from the following observation. When G_1 has no edges coincident with any G_2-edges, the corresponding term is zero since $\mathrm{E}(w_{G_1} w_{G_2}) = \mathrm{E}(w_{G_1})\ \mathrm{E}(w_{G_2})$, due to independence. If there is a single edge in $G = G_1 \bigcup G_2$, the corresponding term is also zero. There are two cases in which the terms in (2.8) may be non-zero. In the first, both G_1 and G_2 have no single edges in themselves and G_1 has at least one edge coincident with an edge of G_2. In the second, there is at least one cycle in both G_1 and G_2. In both cases the number of non-coincident vertices of G is at most h.

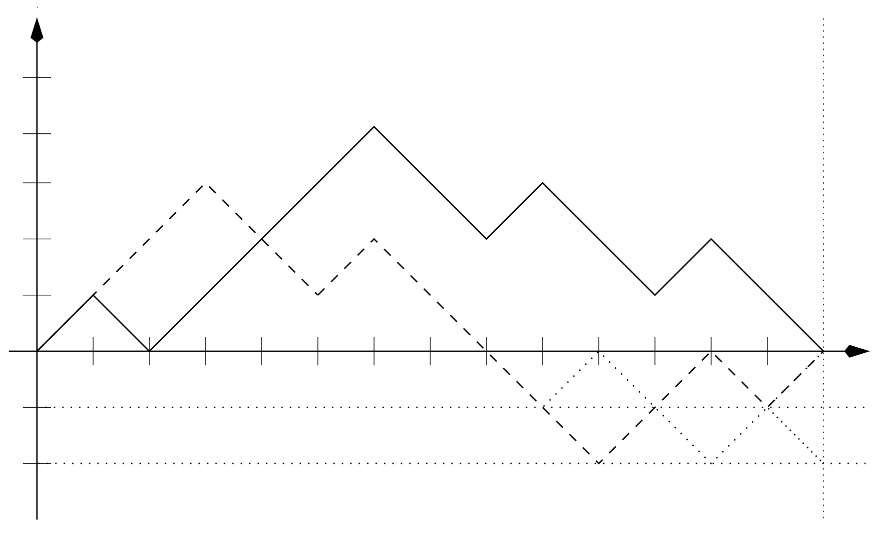

Figure 3. The solid curve represents a CS, the broken curve represents a non-CS and the dotted curve is the reflection of the broken curve.

Remark 2.1. The existence of the second moment of the off-diagonal entries is obviously a necessary and sufficient condition for the semicircular law since the LSD involves the parameter σ^2. It is interesting that there is no moment requirement on the diagonal elements. This fact makes the proof of Theorem 2.12 much easier than exists in the literature.

Sometimes it is of practical interest to consider the case where, for each n, the entries above the diagonal of $\mathbf{W}$ are independent complex random variables with mean zero and variance σ^2, but which may not be identically distributed and may depend on n. We have the following result.

Theorem 2.4. *If* $\mathrm{E}(w_{jk}^{(n)}) = 0$, $\mathrm{E}|w_{jk}^{(n)}|^2 = \sigma^2$ *and for any* $\delta > 0$

$$\lim_{n\to\infty} \frac{1}{\delta^2 n^2} \sum_{jk} \mathrm{E}|w_{jk}^{(n)}|^2 I_{(|w_{jk}^{(n)}| \geq \delta\sqrt{n})} = 0, \tag{2.9}$$

then the conclusion of Theorem 2.1 *holds.*

The proof of this theorem is basically the same as that of Theorem 2.1. At first, we note that one can select a sequence $\delta_n \downarrow 0$ such that (2.9) is still true with δ replaced by δ_n. Then one may truncate the variables at $C = \delta_n\sqrt{n}$. For brevity, in the proof of Theorem 2.4, we suppress the dependence on n from entries of $\mathbf{W}_n$. By Lemma 2.2, we have

$$\|F^{\mathbf{W}_n} - F^{\widetilde{\mathbf{W}}_n}\| \leq \frac{1}{n} \sum_{jk} I_{\{|w_{jk}| \geq \delta_n\sqrt{n}\}},$$

where $\widetilde{\mathbf{W}}_n$ is the matrix of truncated variables. By Condition (2.9),

$$\sum_{jk} \mathrm{P}\{|w_{jk}| \geq \delta_n\sqrt{n}\} \leq \frac{1}{\delta_n^2 n} \sum_{jk} \mathrm{E}|w_{jk}|^2 I_{(|w_{jk}| \geq \delta_n\sqrt{n})} = o(n).$$

Applying Hoeffding's inequality to the sum of the $n(n+1)/2$ independent terms of $I_{(|w_{jk}| \geq \delta_n\sqrt{n})}$, we have

$$\mathrm{P}\Big(\sum_{j\leq k} I_{\{|w_{jk}| \geq \delta_n\sqrt{n}\}} \geq \varepsilon n\Big) \leq \mathrm{P}\Big(\Big|\sum_{j\leq k} [I_{\{|w_{jk}| \geq \delta_n\sqrt{n}\}} - \mathrm{P}(|w_{jk}| \geq \delta_n\sqrt{n})]\Big| \geq \frac{1}{3}\varepsilon n\Big)$$
$$\leq 2\exp\Big(-\frac{9^{-1}\varepsilon^2 n^2}{2[\sum_{jk} \mathrm{P}\{|w_{jk}| \geq \delta_n\sqrt{n}\} + 3^{-1}\varepsilon n]}\Big) \leq 2\exp(-bn),$$

for some positive constant b. By the Borel-Cantelli Lemma, with probability 1, the truncation does not affect the LSD of $\mathbf{W}_n$. Then, applying Lemma 2.3, one can re-centralize the truncated variable and replace the diagonal entries by zero without changing the LSD.

Then for the truncated and re-centralized matrix (still denoted by $\mathbf{W}_n$), it can be shown that, by estimates similar to those given in the proof of Theorem 2.1 and corresponding to (2.7),

$$\frac{1}{n}\mathrm{E}(\mathrm{tr}((n^{-1/2}\mathbf{W}_n)^h)) = \begin{cases} \frac{(2s)!\sigma^h}{s!(s+1)!} + o(1), & \text{if } h = 2s \\ o(1), & \text{if } h = 2s+1. \end{cases} \tag{2.10}$$

However, we cannot prove the counterpart for $\mathrm{Var}(\frac{1}{n}\mathrm{tr}((n^{-1/2}\mathbf{W}_n)^h))$ since its order is only $O(\frac{\delta_n^2}{n})$, which implies convergence "in probability", but not "almost surely". In this case, we can consider the fourth moment and prove

$$\begin{aligned} &\mathrm{E}\Big|\frac{1}{n}\mathrm{tr}((n^{-1/2}\mathbf{W}_n)^h) - \mathrm{E}(\frac{1}{n}\mathrm{tr}((n^{-1/2}\mathbf{W}_n)^h))\Big|^4 \\ &= n^{-4-2h} \sum_{G_1,\ldots,G_4} \mathrm{E}\Big[\prod_{i=1}^4 (w_{G_i} - \mathrm{E}(w_{G_i}))\Big] = O(n^{-2}). \end{aligned} \tag{2.11}$$

In fact, if there is one subgraph among $G_1, \ldots, G_4$ which has no edge coincident with any edges of the other three, the corresponding term is zero. Thus, we only need to estimate those terms for the graphs whose every subgraph has at least one edge coincident with an edge of other subgraphs. Then (2.11) can be proved by analyzing various cases. The details are omitted.

Remark 2.2. In Girko's book (1990), it is stated that condition (2.9) is necessary and sufficient for the conclusion of Theorem 2.4.

2.1.2. Sample covariance matrix

Suppose that $\{x_{jk},\ j,k = 1,2,\ldots\}$ is a double array of i.i.d. complex random variables with mean zero and variance σ^2. Write $\mathbf{x}_k = (x_{1k},\ldots,x_{pk})'$ and $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$. The sample covariance matrix is usually defined by $\mathbf{S} = \frac{1}{n-1}\sum_{k=1}^n(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^*$. However, in spectral analysis of LDRM, the sample covariance matrix is simply defined as $\mathbf{S} = \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^* = \frac{1}{n}\mathbf{X}\mathbf{X}^*$.

The first success in finding the LSD of $\mathbf{S}$ is due to Marčenko and Pastur (1967). Subsequent work was done in Bai and Yin (1988a), Grenander and Silverstein (1977), Jonsson (1982), Wachter (1978) and Yin (1986). When the entries of $\mathbf{X}$ are not independent, Yin and Krishnaiah (1985) investigated the LSD of $\mathbf{S}$ when the underlying distribution is isotropic. The next theorem is a consequence of a result in Yin (1986), where the real case is considered. Here we state it in the complex case.

Theorem 2.5. *Suppose that $p/n \to y \in (0,\infty)$. Under the assumptions stated at the beginning of this section, the ESD of $\mathbf{S}$ tends to a limiting distribution with* *density*

$$p_y(x) = \begin{cases} \frac{1}{2\pi xy\sigma^2}\sqrt{(b-x)(x-a)}, & \text{if } a \le x \le b \\ 0, & \text{otherwise,} \end{cases} \tag{2.12}$$

and a point mass $1 - 1/y$ *at the origin if* $y > 1$, *where* $a = a(y) = \sigma^2(1 - y^{1/2})^2$ *and* $b = b(y) = \sigma^2(1 + y^{1/2})^2$.

The limit distribution of Theorem 2.5 is called the Marčenko — Pastur law with ratio index y and scale index σ^2. The proof relies on the following lemmas.

Lemma 2.6 (Rank Inequality). *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two* $p \times n$ *complex matrices. Then*

$$||F^{\mathbf{AA}^*} - F^{\mathbf{BB}^*}|| \le \frac{1}{p}\text{rank}(\mathbf{A} - \mathbf{B}). \tag{2.13}$$

From Lemma 2.2, one easily derives a weaker result (but enough for applications to LSA of large sample covariance matrices) that $||F^{\mathbf{AA}^*} - F^{\mathbf{BB}^*}|| \le \frac{2}{p}\text{rank}(\mathbf{A} - \mathbf{B})$.

To prove Lemma 2.6, one may assume that $\mathbf{A}' = (\mathbf{A}_1' \vdots \mathbf{A}_2')'$ and $\mathbf{B}' = (\mathbf{B}_1' \vdots \mathbf{A}_2')'$, where the number of rows of $\mathbf{A}_1$ (also $\mathbf{B}_1$) is k =rank$(\mathbf{A}-\mathbf{B})$. Then, as in the proof of Lemma 2.2, Lemma 2.6 can be proven by applying the interlacing inequality to the matrices $\mathbf{A}_2\mathbf{A}_2^*$, $\mathbf{AA}^*$ and $\mathbf{BB}^*$.

Lemma 2.7 (Difference Inequality). *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two* $p \times n$ *complex matrices. Then*

$$L^4(F^{(\mathbf{AA}^*)}, F^{(\mathbf{BB}^*)}) \le \frac{2}{p^2}\text{tr}((\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^*)\text{tr}(\mathbf{AA}^* + \mathbf{BB}^*). \tag{2.14}$$

This lemma relies on the following:

$$L^4(F^{(\mathbf{AA}^*)}, F^{(\mathbf{BB}^*)}) \le \Big(\frac{1}{p}\sum_{k=1}^p |\lambda_k - \eta_k|\Big)^2 \le \frac{2}{p^2}\Big(\sum_{k=1}^p (\sqrt{\lambda_k} - \sqrt{\eta_k})^2\Big)\Big(\sum_{k=1}^p (\lambda_k + \eta_k)\Big),$$

$$\text{tr}(\mathbf{AA}^*) = \sum_{k=1}^p \lambda_k, \qquad \text{tr}(\mathbf{BB}^*) = \sum_{k=1}^p \eta_k,$$

and for some unitary matrices $\mathbf{U} = (u_{jk})$ and $\mathbf{V} = (v_{jk})$,

$$\text{Re}(\text{tr}(\mathbf{AB}^*)) = \sum_{j,k} \sqrt{\lambda_j \eta_k}\text{Re}(u_{jk}\bar{v}_{jk})$$

$$\le \Big(\sum_{j,k} \sqrt{\lambda_j \eta_k}|u_{jk}|^2 \sum_{j,k} \sqrt{\lambda_j \eta_k}|v_{jk}|^2\Big)^{1/2} \le \sum_{k=1}^p \sqrt{\lambda_k \eta_k}.$$

Now we are in a position to sketch a proof of Theorem 2.5. Define $\tilde{x}_{jk} = x_{jk}I_{(|x_{jk}|<C)} - \text{E}(x_{jk}I_{(|x_{jk}|<C)})$ and denote by $\widetilde{\mathbf{S}}$ the sample covariance matrix

constructed of the $\tilde{x}_{jk}$. Similar to the proof of (2.6), employing Lemma 2.7, we can show that with probability 1

$$\limsup_{n\to\infty} L^4(F^{\mathbf{S}}, F^{\widetilde{\mathbf{S}}}) \le 4\sigma^2 \mathrm{E}|x_{11}^2| I_{(|x_{11}|\ge C)}.$$

Also, $\mathrm{E}(|\tilde{x}_{jk}|^2) \to \sigma^2$ as $C \to \infty$. Therefore, in the proof of Theorem 2.5, we may assume that the variables x_{jk} are uniformly bounded, since the right hand side in the above inequality can be arbitrarily small if C is chosen large enough. Then we use the expression

$$\begin{aligned} p^{-1}\mathrm{tr}(\mathbf{S}^h) &= p^{-1}n^{-h} \sum_{i_1,\ldots,i_h} \sum_{j_1,\ldots,j_h} x_{i_1j_1}\bar{x}_{i_2j_1}x_{i_2j_2}\bar{x}_{i_3j_2}\cdots \bar{x}_{i_hj_{h-1}}x_{i_hj_h}\bar{x}_{i_1j_h} \\ &:= p^{-1}n^{-h}\sum_G x_G, \end{aligned}$$

where in the summations, the indices $i_1,\ldots,i_h$ run over $1,\ldots,p$, the indices $j_1,\ldots,j_h$ run over $1,\ldots,n$. An S-graph G is constructed by plotting the i's and j's on two parallel straight lines respectively, drawing h (down) edges from i_v to j_v and another h (up) edges from j_v to i_{v+1} (with the convention $i_{h+1} = i_1$), $v = 1,\ldots,h$. Finally, we show that

$$\mathrm{E}(p^{-1}\mathrm{tr}(\mathbf{S}^h)) = p^{-1}n^{-h}\sum_G \mathrm{E}(x_G) = \sigma^{2h}\sum_{r=0}^{h-1}\frac{y_n^r}{r+1}\binom{h}{r}\binom{h-1}{r} + O(n^{-1}) \quad (2.15)$$

where $y_n = p/n$, and

$$\mathrm{Var}(p^{-1}\mathrm{tr}(\mathbf{S}^h)) = p^{-2}n^{-2h}\sum_{G_1,G_2}[\mathrm{E}(x_{G_1}x_{G_2}) - \mathrm{E}(x_{G_1})\mathrm{E}(x_{G_2})] = O(n^{-2}). \quad (2.16)$$

Similar to the proof of (2.7), the proof of (2.15) reduces to, for each $r = 0,\ldots,h-1$, the calculation of the number of graphs which have no single edges, $r+1$ non-coincident i-vertices and $h-r$ noncoincident j-vertices. In such graphs, each down edge (a,b) must coincide with and only with the up-edge (b,a) which contributes a factor $\mathrm{E}|x_{ab}|^2 \to \sigma^2$ (as $C \to \infty$). We say that two S-graphs are *isomorphic* if one can be coverted to the other through a permutation of $\{1,\ldots,p\}$ on the i-line and a permutation of $\{1,\ldots,n\}$ on the j-line. To compute the number of isomorphic classes, define $d_\ell = -1$ if the path of the graph ultimately leaves an i-vertex (other than the initial i-vertex) after the ℓth down edge and define $u_\ell = 1$ if the ℓth up-edge leads to a new i-vertex. For other cases, define d_ℓ and u_ℓ as 0. Note that we always have $d_1 = 0$. It is obvious that $u_1 + \cdots + u_{\ell-1} + d_1 + \cdots + d_\ell \ge 0$. Ignoring this restriction, we have $\binom{h}{r}\binom{h-1}{r}$ ways to arrange r ones into the h positions of up-edges and r minus ones into the

$h-1$ positions of down-edges (except the first). If ℓ is the first integer such that $u_1+\cdots+u_{\ell-1}+d_1+\cdots+d_\ell<0$, then $u_{\ell-1}=0$ and $d_\ell=-1$. By changing $u_{\ell-1}$ to 1 and d_ℓ to 0, we get a d-sequence with $r-1$ minus ones and a u-sequence with $r+1$ ones. Thus, the number of isomorphic classes is $\binom{h}{r}\binom{h-1}{r}-\binom{h}{r+1}\binom{h-1}{r-1}=\frac{1}{r+1}\binom{h}{r}\binom{h-1}{r}$. The number of graphs in each isomorphic class is obviously $p(p-1)\cdots(p-r)n(n-1)\cdots(n-h+r-1)=pn^h y_n^r(1+O(n^{-1}))$. Then (2.15) follows. The proof of (2.16) is similar to that of (2.8). This completes the proof of the theorem.

Remark 2.3. The existence of the second moment of the entries is obviously necessary and sufficient for the Marčenko-Pastur Law since the LSD involves the parameter σ^2. The condition of zero mean can be relaxed to having a common mean, since the means of the entries form a rank-one matrix which can be removed by applying Lemma 2.6.

Remark 2.4. As argued before the statement of Theorem 2.4, sometimes it is of practical interest to consider the case where the entries of $\mathbf{X}$ depend on n, and for each n they are independent but not identically distributed. Similar to Theorem 2.4, truncating the variables at $\delta_n\sqrt{n}$ for some $\delta_n\downarrow 0$ by using Lemma 2.6, and recentralizing by using Lemma 2.7, one can prove the following generalization.

Theorem 2.8. *Suppose that for each n, the entries of $\mathbf{X}_n$ are independent complex variables, with a common mean and variance σ^2. Assume that $p/n\to y\in(0,\infty)$ and that for any $\delta>0$,*

$$\frac{1}{\delta^2 np}\sum_{j,k}\mathrm{E}(|x_{jk}^{(n)}|^2 I_{(|x_{jk}^{(n)}|\ge\delta\sqrt{n})})\to 0. \tag{2.17}$$

Then $F^{\mathbf{S}}$ tends almost surely to the Marčenko-Pastur law with ratio index y and scale index σ^2.

Now we consider the case $p\to\infty$ but $p/n\to 0$ as $p\to\infty$. It is conceivable that almost all eigenvalues tend to 1 and hence the ESD of $\mathbf{S}$ tends to a degenerate one. In turn, to investigate the behavior of the eigenvalues of the sample covariance matrix $\mathbf{S}$, we consider the ESD of the matrix $\mathbf{W}=\sqrt{n/p}(\mathbf{S}-\sigma^2\mathbf{I}_p)=\frac{1}{\sqrt{np}}(\mathbf{X}\mathbf{X}^*-n\sigma^2\mathbf{I}_p)$. When the entries of $\mathbf{X}$ are real, under the existence of the fourth moment, Bai and Yin (1988a) showed that its ESD tends to the semicircular law almost surely as $p\to\infty$. Now we give a generalization of this result.

Theorem 2.9. *Suppose that for each n the entries of the matrix $\mathbf{X}_n$ are independent complex random variables with a common mean and variance σ^2. Assume that for any constant $\delta>0$, as $p\to\infty$ with $p/n\to 0$,*

$$\frac{1}{p\delta^2\sqrt{np}}\sum_{jk}\mathrm{E}(|x_{jk}^{(n)}|^2 I_{(|x_{jk}^{(n)}|\ge\delta\sqrt[4]{np})})=o(1) \tag{2.18}$$

and

$$\frac{1}{np^2}\sum_{jk} \mathrm{E}(|x_{jk}^{(n)}|^4 I_{(|x_{jk}^{(n)}|\le \delta \sqrt[4]{np})}) = o(1). \tag{2.19}$$

Then with probability 1 *the ESD of* $\mathbf{W}$ *tends to the semi-circular law with scale index* σ^2.

Remark 2.5. Conditions (2.18) and (2.19) hold if the entries of $\mathbf{X}$ have bounded fourth moments. This is the condition assumed in Bai and Yin (1988a).

The proof of Theorem 2.9 consists of the following steps. Applying Lemma 2.6, we may assume that the common mean is zero. Truncate the entries of $\mathbf{X}$ at $\delta_p \sqrt[4]{np}$, where $\delta_p \to 0$ such that (2.18) and (2.19) hold with δ replaced by δ_p. By Condition (2.18), $\sum_{jk} \mathrm{P}(|x_{jk}^{(n)}| \ge \delta_p \sqrt[4]{np}) = o(p)$. From this and applying Hoeffding's inequality, one can prove that the probability that the number of truncated elements of $\mathbf{X}$ is greater than εp is less than Ce^{-bp} for some $b > 0$.

One needs to recentralize the truncated entries of $\mathbf{X}$. The application of Lemma 2.7 requires

$$\frac{1}{p}\sum_{jk} |\mathrm{E}(x_{jk}^{(n)}) I_{[|x_{jk}^{(n)}|\ge \delta_p \sqrt[4]{np}]}|^2 = o(1)$$

and

$$\frac{1}{np^2}\sum_{jk} |x_{jk}^{(n)}|^2 I_{[|x_{jk}^{(n)}|\le \delta_p \sqrt[4]{np}]} = o(1), \text{a.s.}$$

Here, the first assertion is an easy consequence of (2.18). The second can be proved by applying Bernstein's inequality (see Prokhorov (1968)).

The next step is to remove the diagonal elements of $\mathbf{W}$. Write $y_\ell = I_{\{\frac{1}{\sqrt{np}}\left|\sum_{j=1}^n (|x_{\ell j}|^2-\sigma^2)\right|\ge\varepsilon\}}$. Note that by Condition (2.19),

$$\sum_{\ell=1}^p \mathrm{E}(y_\ell) \le \frac{1}{\varepsilon^2\sqrt{np}}\sum_{\ell,j} \mathrm{E}(|x_{\ell j}|^4 I_{|x_{\ell j}|\le \delta_p \sqrt[4]{np}}) = o(p).$$

Applying Hoeffding's inequality, we have

$$\mathrm{P}(\sum_{\ell=1}^p y_\ell \ge \varepsilon p) \le 2e^{-bp} \tag{2.20}$$

for some $b > 0$. Then applying Lemma 2.2, we can replace the diagonal elements of $\mathbf{W}$ which are greater than ε by zero, since the number of such elements is $o(p)$ by (2.20). By Lemma 2.3, we can also replace those smaller than ε by zero.

In the remainder of the proof, we assume that $\mathbf{W} = (\frac{1}{\sqrt{np}} \sum_{j=1}^n x_{i_1 j}\overline{x}_{i_2 j}(1-\delta_{i_1,i_2}))$, where δ_{jk} is the Kronecker delta . Then we need to prove that

$$\mathrm{E}(\mathrm{tr}(\mathbf{W}^h)) = \begin{cases} \frac{(2s)!\sigma^{2h}}{s!(s+1)!} + o(1), & \text{if } h = 2s \\ o(1), & \text{if } h = 2s+1 \end{cases}$$

$$\mathrm{E}|\mathrm{tr}(\mathbf{W}^h) - \mathrm{E}(\mathrm{tr}(\mathbf{W}^h))|^4 = O(\frac{1}{p^2}).$$

Similar to the proof of Theorem 2.5, construct graphs for estimating $\mathrm{E}(\mathrm{tr}(\mathbf{W}^h))$. Denote by r and s the numbers of i and j vertices. Note that the number of non-coincident edges is not less than twice the number of non-coincident j vertices, since consecutive i vertices are not equal. It is obvious that the number of non-coincident edges is not less than $r+s-1$. Therefore, the contribution of each isomorphic class to the sum is not more than

$$\begin{cases} p^{-1}(np)^{-h/2}n^s p^r(\delta_p \sqrt[4]{np})^{2h-4s}\sigma^{4s} = \delta_p^{2h-4s}p^{r-s-1}\sigma^{4s} & \text{if } s+1 \geq r, \\ p^{-1}(np)^{-h/2}n^s p^r(\delta_p \sqrt[4]{np})^{2h-2s-2r+2}\sigma^{2s+2r} & \\ \quad = \delta_p^{2h-2s-2r+2}(p/n)^{r-s-1}\sigma^{2s+2r} & \text{if } s+1 < r. \end{cases}$$

The quantities on the right hand sides of the above estimations are $o(1)$ unless $h = 2s = 2r-2$. When $h = 2s = 2r-2$, the contribution of each isomorphic class is $\sigma^{2h}(1+O(p^{-1}))$ and the number of non-isomorphic graphs is $(2s)!/[s!(s+1)!]$. The rest of the proof is similar to that of Theorem 2.4 and hence omitted.

2.1.3. Product of two random matrices

The motivation for studying products of random matrices originates from the investigation of the spectral theory of large sample covariance matrices when the population covariance matrix is not a multiple of an identity matrix, and that of *multivariate* $\mathbf{F} = \mathbf{S}_1\mathbf{S}_2^{-1}$ *matrices*. When $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent Wishart, the LSD of $\mathbf{F}$ follows from Wachter (1980) and its explicit forms can be found in Bai, Yin and Krishnaiah (1987), Silverstein (1985a) and Yin, Bai and Krishnaiah (1983). Relaxation of the Wishart assumption on $\mathbf{S}_1$ and $\mathbf{S}_2$ relies on the investigation of the strong limit of the smallest eigenvalue of a sample covariance matrix. Based on the results in Bai and Yin (1993) and Yin (1986), and using the approach in Bai, Yin and Krishnaiah (1985), one can show that the LSD of $\mathbf{F}$ is the same as if both $\mathbf{S}_1$ and $\mathbf{S}_2$ were Wishart when the underlying distribution of $\mathbf{S}_1$ has finite second moment and that of $\mathbf{S}_2$ has finite fourth moment. Yin and Krishnaiah (1983) investigated the limiting distribution of a product of a Wishart matrix $\mathbf{S}$ and a positive definite matrix $\mathbf{T}$. Later work was done in Bai, Yin and Krishnaiah (1986), Silverstein (1995), Silverstein and Bai (1995) and Yin (1986). Here we present the following result.

Theorem 2.10. *Suppose that the entries of* $\mathbf{X}$ *are independent complex random variables satisfying* (2.17)*, and assume that* $\mathbf{T}(=\mathbf{T}_n)$ *is a sequence of* $p \times p$ *Hermitian matrices independent of* $\mathbf{X}$ *such that its ESD tends to a non-random and non-degenerate distribution* H *in probability* (or almost surely). *Further assume that* $p/n \to y \in (0, \infty)$. *Then the ESD of the product* $\mathbf{ST}$ *tends to a non-random limit in probability* (*or almost surely, respectively*).

This theorem contains Yin (1986) as a special case. In Yin (1986), the entries of $\mathbf{X}$ are assumed to be real and i.i.d. with mean zero and variance 1, the matrix $\mathbf{T}$ is assumed to be real and positive definite and to satisfy, for each fixed h,

$$\frac{1}{p}\mathrm{tr}(\mathbf{T}^h) \to \alpha_h, \quad \text{(in pr. or a.s.,)} \tag{2.21}$$

where the sequence $\{\alpha_h\}$ satisfies Carleman's condition.

There are two directions to generalize Theorem 2.10. One is to relax the independence assumption on the entries of $\mathbf{S}$. Bai, Yin and Krishnaiah (1986) assume the columns of $\mathbf{X}$ are i.i.d. and each column is isotropically distributed with certain moment conditions, for example. It could be that Theorems 2.1, 2.4, 2.5, 2.8 and 2.10 are still true when the underlying variables defining the Wigner or sample covariance matrices are weakly dependent, say ϕ-mixing, although I have not found any such results yet. It may be more interesting to investigate the case where the entries are dependent, say the columns of $\mathbf{X}$ are i.i.d. and the entries of each column are uncorrelated but not independent.

Another direction is to generalize the structure of the setup. An example is given in Theorem 3.4. Since the original proof employs the Stieltjes transformation technique, we postpone its statement and proof to Section 3.1.2.

To sketch the proof of Theorem 2.10, we need the following lemma.

Lemma 2.11. *Let* G^0 *be a connected graph with* m *vertices and* h *edges. To each vertex* $v(=1, \ldots, m)$ *there corresponds a positive integer* n_v*, and to each edge* $e_j = (v_1, v_2)$ *there corresponds a matrix* $\mathbf{T}_j = (t_{\eta,\tau}^{(j)})$ *of order* $n_{v_1} \times n_{v_2}$. *Let* E_c *and* E_{nc} *denote the sets of cutting edges* (*those edges whose removal causes the graph disconnected*) *and non-cutting edges, respectively. Then there is a constant* C*, depending upon* m *and* h *only, such that*

$$\Big|\sum_{i_1,\ldots,i_m}\prod_{j=1}^{h} t^{(j)}_{i_{f_{ini}(e_j)},i_{f_{end}(e_j)}}\Big| \le Cn \prod_{e_j \in E_{nc}} \|\mathbf{T}_j\| \prod_{e_j \in E_c} \|\mathbf{T}_j\|_0,$$

where $n = \max(n_1, \ldots, n_m)$*,* $\|\mathbf{T}_j\|$ *denotes the maximum singular value, and* $\|\mathbf{T}_j\|_0$ *equals the product of the maximum dimension and the maximum absolute* *value of the entries of* $\mathbf{T}_j$*; in the summation* i_v *runs over* $\{1, \ldots, n_v\}$*,* $f_{ini}(e_j)$ *and* $f_{end}(e_j)$ *denote the initial and end vertices of the edge* e_j*.*

If there are no cutting edges in G^0, then the lemma follows from the norm inequality $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$. For the general case, the lemma can be proved by induction with respect to the number of cutting edges. The details are omitted.

In the proof of Theorem 2.10, we only consider a.s. convergence, since the case of convergence in probability can be reduced to the a.s. case by using the strong representation theorem (see Yin (1986) for details).

For given $\tau_0 > 0$, define a matrix $\widetilde{\mathbf{T}}$ by replacing, in the spectral decomposition of $\mathbf{T}$, the eigenvalues of $\mathbf{T}$ whose absolute values are greater than τ_0 by zero. Then the ESD of $\widetilde{\mathbf{T}}$ converges to

$$H_{\tau_0}(x) = \int_{-\infty}^{x} I_{[-\tau_0,\tau_0]}(u)H(du) + (H(-\tau_0) + 1 - H(\tau_0))I_{[0,\infty)}(x), \quad \text{a.s.}$$

and (2.21) holds, with $\tilde{\alpha}_h = \int_{|x|\leq\tau_0} x^h dH(x)$. An application of Lemma 2.2 shows that the substitution of $\mathbf{T}$ by $\widetilde{\mathbf{T}}$ alters the ESD of the product by at most $\frac{1}{p}\#\{j : |\lambda_j(\mathbf{T})| \geq \tau_0\}$, which can be arbitrarily uniformly small by choosing τ_0 large.

We claim that Theorem 2.10 follows if we can prove that, with probability 1, $F^{\mathbf{S}\widetilde{\mathbf{T}}}$ converges to a non-degenerate distribution F_{τ_0} for each fixed τ_0. First, we can show the tightness of $\{F^{\mathbf{ST}}\}$ from $F^{\mathbf{T}} \to H$ and the inequality

$$\begin{aligned} F^{\mathbf{ST}}(M) - F^{\mathbf{ST}}(-M) &\geq F^{\mathbf{S}\widetilde{\mathbf{T}}}(M) - F^{\mathbf{S}\widetilde{\mathbf{T}}}(-M) - 2||F^{\mathbf{ST}} - F^{\mathbf{S}\widetilde{\mathbf{T}}}|| \\ &\geq F^{\mathbf{S}\widetilde{\mathbf{T}}}(M) - F^{\mathbf{S}\widetilde{\mathbf{T}}}(-M) - 2(F^{\mathbf{T}}(-\tau_0) + 1 - F^{\mathbf{T}}(\tau_0)). \end{aligned}$$

Here, the second inequality follows by using Lemma 2.2. Second, we can show that any convergent subsequences of $\{F^{\mathbf{ST}}\}$ have the same limit by using the inequality

$$|F_1(x) - F_2(x)| \leq \sum_{j=1}^{2} \Big[|F_j(x) - F_{n_j}^{\mathbf{ST}}(x)| + ||F_{n_j}^{\mathbf{ST}} - F_{n_j}^{\mathbf{S}\widetilde{\mathbf{T}}}||\Big] + |F_{n_1}^{\mathbf{S}\widetilde{\mathbf{T}}}(x) - F_{n_2}^{\mathbf{S}\widetilde{\mathbf{T}}}(x)|,$$

where F_1 and F_2 denote the limits of two convergence subsequences $\{F_{n_1}^{\mathbf{ST}}\}$ and $\{F_{n_2}^{\mathbf{ST}}\}$ respectively. This completes the proof of the assertion.

Consequently, the proof of Theorem 2.10 reduces to showing that $\{F^{\mathbf{S}\widetilde{\mathbf{T}}}\}$ converge to a non-random limit. Again, using Lemma 2.2, we may assume that the entries of $\mathbf{X}$ are truncated at $\sqrt{n}\delta_n$ ($\delta_n \to 0$) and centralized. In the sequel, for convenience, we still use $\mathbf{X}$ and $\mathbf{T}$ to denote the truncated matrices.

After truncation and centralization, one can see that $|x_{jk}| \le \delta_n\sqrt{n}$ and

$$\mathrm{E}|x_{jk}|^2 \le \sigma^2 \quad \text{with} \quad \frac{1}{np}\sum_{jk}\mathrm{E}|x_{jk}|^2 \to \sigma^2. \tag{2.22}$$

To estimate the moments of the ESD, we have

$$\frac{1}{p}\mathrm{E}[(\mathbf{ST})^h|\mathbf{T}] = p^{-1}n^{-h}\sum_G \mathrm{E}((tx)_G|\mathbf{T}), \tag{2.23}$$

where

$$(tx)_G = x_{i_1j_1}\overline{x}_{i_2j_1}t_{i_2i_3}x_{i_3j_2}\overline{x}_{i_4j_2}\cdots x_{i_{2h-1}j_h}\overline{x}_{i_{2h}j_h}t_{i_{2h}i_1}.$$

The Q-graphs (named in Yin and Krishnaiah (1983)) are drawn as follows: as before, plot the vertices i's and j's on two parallel lines and draw h (down) edges from i_{2u-1} to j_u, h (up) edges from j_u to i_{2u} and h (horizontal) edges from i_{2u} to i_{2u+1} (with the convention $i_{2h+1} = i_1$). If there is a single vertical edge in G, then the corresponding term is zero. We split the sum of non-zero terms in (2.23) into subsums in accordance with isomorphic classes of graphs without single vertical edges. For a Q-graph G in a given isomorphic class, denote by s the number of non-coincident j-vertices and by r the number of disjoint connected blocks of horizontal edges. Glue the coincident vertical edges and denote the resulting graph by G^0. Let the $p\times p$-matrix $\mathbf{T}$ correspond to each horizontal edge of G^0 and let the $p\times n$-matrix $\mathbf{T}^{(x)}_{\mu,\nu} = (\mathrm{E}(x^{\mu}_{i,j}\overline{x}^{\nu}_{i,j}))$ correspond to each vertical edge of G^0 that consists of μ down edges and ν up edges of G. Note that $\mu+\nu\ge 2$ and $|\mathrm{E}(x^{\mu}_{i,j}\overline{x}^{\nu}_{i,j})| \le \sigma^2(\delta_n\sqrt{n})^{\mu+\nu-2}$. It is obvious that $\|\mathbf{T}\|\le\tau_0$, $\|\mathbf{T}^{(x)}_{\mu,\nu}\| \le \sqrt{np}\sigma^2(\delta_n\sqrt{n})^{\mu+\nu-2}$ and $\|\mathbf{T}^{(x)}_{\mu,\nu}\|_0 \le \max(n,p)\sigma^2(\delta_n\sqrt{n})^{\mu+\nu-2}$. Also, every horizontal edge of G^0 is non-cutting. Split the right hand side of (2.23) as J_1+J_2 where J_1 corresponds to the sum of those terms whose graphs G^0 contain at least one vertical edge with multiplicity greater than 2 and J_2 is the sum of all other terms. Applying Lemma 2.11, we get $J_1 = O(\delta_n^2) = o(1)$.

We further split J_2 as $J_{21}+J_{22}$, where J_{21} is the sum of all those terms whose G_0-graphs contain at least one non-cutting vertical edge and J_{22} is the sum of the rest. For graphs corresponding to the terms in J_{21}, we must have $s+r\le h$. When evaluating J_{21}, we fix the indices $j_1,\ldots,j_s$ and perform the summation for $i_1,\ldots,i_r$ first. Corresponding to the summation for fixed $j_1,\ldots,j_s$, we define a new graph $G(j_1,\ldots,j_s)$ as follows: If (i_g,j_h) is a vertical edge of G_0, consisting of μ-up and ν-down edges of G (note that $\mu+\nu=2$), then remove this edge and add to the vertex i_g a loop, to which there corresponds the $p\times p$ diagonal matrix $\mathbf{T}(j_h) = \mathrm{diag}(E(x^{\mu}_{1,j_h}\bar{x}^{\nu}_{1,j_h}),\ldots,E(x^{\mu}_{p,j_h}\bar{x}^{\nu}_{p,j_h}))$, see Figure 4. After all vertical edges of G^0 are removed, the r disjoint connected blocks of the resulting graph $G(j_1,\ldots,j_s)$ have no cutting edges. Note that the $\|\cdot\|$-norms of the diagonal matrices are not greater than σ^2. Applying Lemma 2.11 to each of connected blocks, we obtain

$$|J_{21}| \leq Cp^{-1}n^{-h} \sum_{s+r\leq h} \sum_{j_1,\ldots,j_s} p^r \sigma^{2h} \tau_0^h = O(1/n).$$

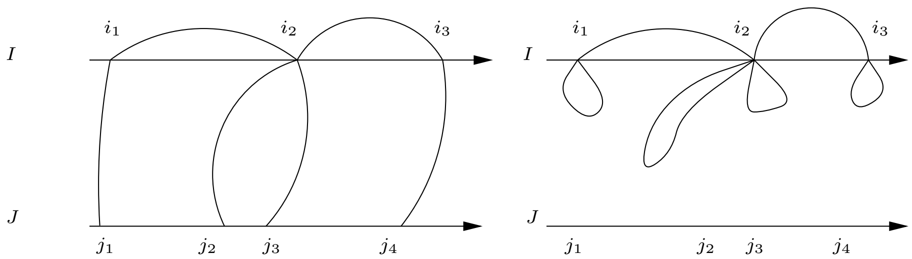

Figure 4. The left graph is the original one and the right one is the resulting graph.

Finally, consider J_{22}. Since all vertical edges are cutting edges, we have $s + r = h + 1$. There are exactly h non-coincident vertical edges, in which each down-edge (a, b) must coincide with one and only one up-edge (b, a). Thus, the contribution of the expectations of the x-variables is $\prod_{\ell=1}^h \mathrm{E}(|x^2_{i_{f_{ini}(e_\ell)}, j_{f_{end}(e_\ell)}}|)$. For a given vertical edge, if its corresponding matrix $\mathbf{T}^{(x)}$ is replaced by the $p \times n$ matrix of all entries σ^2, applying Lemma 2.11 again, this will cause a difference of $o(1)$ in J_{22}, since the norms ($\|\cdot\|$ or $\|\cdot\|_0$) of the difference matrix are only $o(n)$, by (2.22).

Now, denote by $\mu_1, \ldots, \mu_r$ the sizes (the numbers of edges) of the r disjoint blocks of horizontal edges. Then it is not difficult to show that for each class of isomorphic graphs, the sub-sum in J_{22} tends to $y^{r-1}\alpha_{\mu_1} \cdots \alpha_{\mu_r}(1 + o(1))$. Thus, to evaluate the right hand side of (2.23), one only needs to count the number of isomorphic classes.

Let i_m denote the number of disjoint blocks of horizontal subgraphs of size m. Then it can be shown that the number of isomorphic classes is $\frac{h!}{s!i_1!\cdots i_s!}$. For details, see Yin (1986). Hence,

$$\begin{aligned}
\frac{1}{p}\mathrm{E}[(\mathbf{ST})^h|\mathbf{T}] &= \sigma^{2h} \sum_{s=1}^h y^{h-s} \sum \frac{h!}{s!} \prod_{m=1}^s \left(\frac{[p^{-1}\mathrm{tr}(\mathbf{T}^m)]^{i_m}}{i_m!}\right) + o(1) \\
&= \sigma^{2h} \sum_{s=1}^h y^{h-s} \sum \frac{h!}{s!} \prod_{m=1}^s \frac{\alpha_m^{i_m}}{i_m!} + o(1), \qquad (2.24)
\end{aligned}$$

where the inner summation is taken with respect to all nonnegative integer solutions of $i_1 + \cdots + i_s = h + 1 - s$ and $i_1 + 2i_2 + \cdots + si_s = h$.

Similar to the proof of (2.11), to complete the proof of the theorem, one needs to show that

$$\mathrm{E}(|\frac{1}{p}(\mathbf{ST})^h - \mathrm{E}(\frac{1}{p}(\mathbf{ST})^h|\mathbf{T}]|^4|\mathbf{T}] = O(n^{-2}),$$

whose proof is similar to, and easier than, that of (2.24). The convergence of the ESD of **ST** and the non-randomness of the limiting distribution then follow by verifying Carleman's condition.

2.2. Limits of extreme eigenvalues

In multivariate analysis, many statistics involved with a random matrix can be written as functions of integrals with respect to the ESD of the random matrix. When applying the Helly-Bray theorem to find an approximate value of the statistics, one faces the difficulty of dealing with integrals with unbounded integrands. Thus, the study of strong limits of extreme eigenvalues is an important topic in spectral analysis of LDRM.

2.2.1. Limits of extreme eigenvalues of the Wigner matrix

The following theorem is a generalization of Bai and Yin (1988b) where the real case is considered. The complex case is treated here because the question often arises as to whether the result is true for the complex case.

Theorem 2.12. *Suppose that the diagonal elements of the Wigner matrix* $\mathbf{W}$ *are i.i.d. real random variables, the elements above the diagonal are i.i.d. complex random variables, and all these variables are independent. Then the largest eigenvalue of* $n^{-1/2}\mathbf{W}$ *tends to* $2\sigma > 0$ *with probability* 1 *if and only if the following four conditions are true.*

$$\begin{aligned}
&\text{(i) } \mathrm{E}((w_{11}^+)^2) < \infty,\\
&\text{(ii) } \mathrm{E}(w_{12}) \text{ is real and } \le 0,\\
&\text{(iii) } \mathrm{E}(|w_{12} - \mathrm{E}(w_{12}|^2) = \sigma^2,\\
&\text{(iv) } \mathrm{E}(|w_{12}^4|) < \infty,
\end{aligned} \tag{2.25}$$

where $x^+ = \max(x, 0)$.

The proof of the sufficiency part of Theorem 2.12 consists of the following steps. First, by Theorem 2.1, we have

$$\liminf_{n\to\infty} \lambda_{\max}(n^{-1/2}\mathbf{W}) \ge 2\sigma, \text{ a.s.} \tag{2.26}$$

Thus, the problem reduces to proving

$$\limsup_{n\to\infty} \lambda_{\max}(n^{-1/2}\mathbf{W}) \le 2\sigma, \text{ a.s.} \tag{2.27}$$

Let $\widetilde{\mathbf{W}}$ be the matrix obtained from $\mathbf{W}$ by replacing the diagonal elements with zero and centralizing the off diagonal elements. By Conditions (i) and (ii), we notice that $\limsup \frac{1}{\sqrt{n}} w_{kk}^+ = 0$, a.s. Then

$$\lambda_{\max}(n^{-1/2}\mathbf{W}) = n^{-1/2} \max_{||\mathbf{x}||=1} \Big(\sum_{j,k} x_j \bar{x}_k w_{jk} \Big)$$
$$\leq \max_{||\mathbf{x}||=1} \Big(\frac{1}{\sqrt{n}} \sum_{j\neq k} x_j \bar{x}_k (w_{jk} - \mathrm{E}(w_{jk})) + \frac{1}{\sqrt{n}} \max_k (w_{kk}^+ - \mathrm{E}(w_{12})) \Big)$$
$$\leq \lambda_{\max}(\widetilde{\mathbf{W}}) + o(1). \tag{2.28}$$

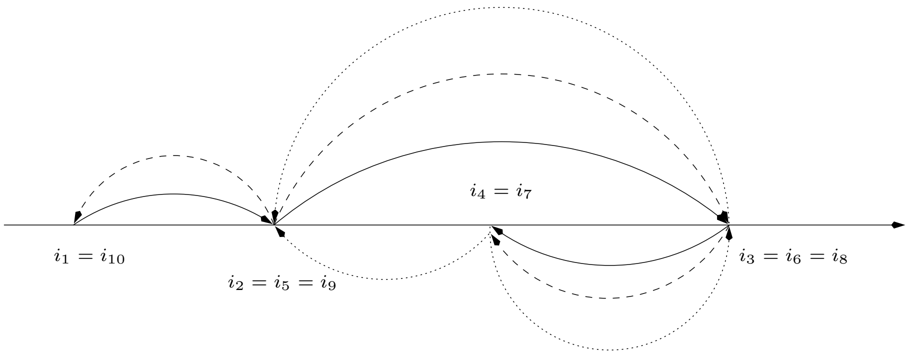

Figure 5. Four types of edges in a W-graph.

This further reduces the proof of (2.27) to showing that $\limsup_{n\to\infty} \lambda_{\max}(\widetilde{\mathbf{W}}) \leq 2\sigma$, a.s.

For brevity of notation, we again use $\mathbf{W}$ for $\widetilde{\mathbf{W}}$, i.e., we assume that the diagonal elements and the means of off diagonal elements of $\mathbf{W}$ are zero. Then by condition (iv), we may select a sequence of constants $\delta_n \to 0$ such that

$$\mathrm{P}(\mathbf{W} \neq \widetilde{\mathbf{W}}, \text{ i.o.}) = 0,$$

where $\widetilde{\mathbf{W}}$ is redefined as $(w_{jk} I_{(|w_{jk}|\leq \delta_n \sqrt{n})})$.

Note that $\mathrm{E}(w_{12}) = 0$ implies

$$\lambda_{\max}(n^{-1/2}\mathrm{E}(\widetilde{\mathbf{W}})) \leq (1 + n^{-1/2}) |\mathrm{E}(w_{12} I_{(|w_{12}|\leq \delta_n \sqrt{n})})| \to 0. \tag{2.29}$$

Therefore, we only need to consider the upper limit of the largest eigenvalue of $\widetilde{\mathbf{W}} - \mathrm{E}(\widetilde{\mathbf{W}})$. For simplicity, we still use $\mathbf{W}$ to denote the truncated and recentralized matrix.

Select a sequence of even integers $h = h_n = 2s$ with the properties $h/\log n \to \infty$ and $h\delta_n^{1/4}/\log n \to 0$. We shall estimate

$$\mathrm{E}(\mathrm{tr}(\mathbf{W}^h)) = \sum_{i_1,\ldots,i_h} \mathrm{E}(w_{i_1 i_2} w_{i_2 i_3} \cdots w_{i_h i_1}) = \sum_G \mathrm{E}(w_G), \tag{2.30}$$

where the graphs are constructed as in the proof of Theorem 2.1. Classify the edges into several types. An edge (i_a, i_{a+1}) is called an innovation or a Type 1 (T_1) edge if $i_{a+1} \notin \{i_1, \ldots, i_a\}$ and is called a Type 3 (T_3) edge if it coincides with an innovation which is single up to i_a. A T_3-edge (i_a, i_{a+1}) is called irregular if there is only one innovation which is single up to i_a and has a vertex coinciding with i_a in the chain $(i_1, \ldots, i_a)$. Otherwise, it is said to be regular. All other edges are called Type 4 (T_4) edges. A T_4-edge is also called a Type 2 (T_2) edge if it does not coincide with any edges prior to it. Examples of the four types of edges are given in Figure 5, in which the first three edges (in solid arcs) are innovations, the broken arcs are T_3 edges and the dotted arcs are T_4 edges. Among the T_3 edges, (i_5, i_6) is a regular T_3 edge (since the path may go to $i_6 = i_1$ instead of $i_6 = i_3$) and among the T_4 edges, (i_4, i_5) is a T_2 edge.

To estimate the right hand side of (2.30), we need the following lemmas whose proofs can be found in Bai and Yin (1988b) or Yin, Bai and Krishnaiah (1988).

Lemma 2.13. *Let t denote the number of non-coincident T_4-edges and u denote the number of innovations which are single up to i_a and have a vertex coinciding with i_a in the chain $(i_1, \ldots, i_a)$. Then $u \le t + 1$.*

Lemma 2.14. *The number of regular T_3-edges is not greater than twice the number of T_4-edges.*

Now, we return to the proof of Theorem 2.12. Suppose that there are r innovations ($r \le s$) and t non-coincident T_4-edges in a graph. Then there are r T_3-edges, $2s - 2r$ T_4-edges and $r + 1$ non-coincident vertices. There are at most n^{r+1} ways to plot the non-coincident vertices and at most $\binom{2s}{r}$ ways to assign the innovations to the $2s$ edges. In a canonical graph (i.e., a graph which starts from the edge $(1, 2)$ and the end-vertex of each innovation is one plus the end-vertex of the previous innovation), there is only one way to plot the innovations. There are at most $\binom{2s-r}{r}$ ways to select the T_3-edges from the remaining $2s - r$ edges and only one way to plot irregular T_3-edges. By Lemmas 2.13 and 2.14, there are at most $(t + 1)^{4(s-r)}$ ways to plot the regular T_3-edges. After the T_1- and T_3-edges have been plotted, there are at most $\binom{2s-2r}{t}$ ways to select the t non-coincident T_4-edges and at most t^{2s-2r} ways to plot the $2s - 2r$ T_4-edges into the t places. Finally, we note that the absolute value of each term is at most $\sigma^{2(r-t)}\mu^t(\sqrt{n}\delta_n)^{2(s-r)-t}$, where $\mu = \mathrm{E}(|w_{12}^3|)$. We obtain

$$\begin{aligned}\mathrm{E}(\mathrm{tr}(\mathbf{W})^h) &\le \sum_{r=1}^{s}\sum_{t=0}^{2s-2r} n^{s+1}\binom{2s}{r}\binom{2s-r}{r}(t+1)^{8(s-r)}\sigma^{2r}(\mu/(\sigma\sqrt{n}\delta_n))^t\delta_n^{2(s-r)}\\ &\le 2sn^{s+1}[2\sigma + \delta_n(8s/\log(\sqrt{n}\delta_n\sigma/\mu))^4]^{2s}\\ &= 2sn^{s+1}[2\sigma + o(1)]^{2s}.\end{aligned}$$

Then by the fact that $\lambda_{\max}^{2s}(n^{-1/2}\mathbf{W}) \le n^{-s}\mathrm{tr}(\mathbf{W}^{2s})$, for any $\varepsilon > 0$, we have

$$\mathrm{P}(\lambda_{\max}(n^{-1/2}\mathbf{W}) \ge 2\sigma + \varepsilon) \le 2sn[1 - \frac{\varepsilon - o(1)}{2\sigma + \varepsilon}]^{2s}.$$

The right hand side above is summable by the choice of s. Therefore, by the Borel Cantelli Lemma, $\limsup \lambda_{\max}(n^{-1/2}\mathbf{W}) \le 2\sigma$ almost surely. The sufficiency is proved.

Conversely, if $\limsup \lambda_{\max}(n^{-1/2}\mathbf{W}) \le 2a$ with $a > 0$, then, by $\lambda_{\max}(n^{-1/2}\mathbf{W}) \ge \max_k n^{-1/2} w_{kk}$, we have $\limsup \max_k n^{-1/2} w_{kk}^{+} \le 2a + \varepsilon$, which implies condition (i).

For $w_{jk} \ne 0$ and $|w_{kk}| \le n^{1/4}$, $|w_{jj}| \le n^{1/4}$, taking $x_k = w_{jk}/(\sqrt{2}|w_{jk}|)$ and $x_j = 1/\sqrt{2}$ in the first equality of (2.28), we have $\lambda_{\max}(n^{-1/2}\mathbf{W}) \ge n^{-1/2}|w_{jk}| - n^{-1/4}$. Thus, $\lambda_{\max}(n^{-1/2}\mathbf{W}) \ge n^{-1/2} \max_{\{|w_{kk}| \le n^{1/4}, |w_{jj}| \le n^{1/4}\}} |w_{jk}| - n^{-1/4}$. This implies Condition (iv), by noticing that $k_n = \#\{k \le n; |w_{kk}| > n^{1/4}\} = o_{a.s.}(n)$.

If $\mathrm{E}(\mathrm{Re}(w_{12})) > 0$, then take $\mathbf{x}$ with kth element $x_k = (n - k_n)^{-1/2}$ or 0 in accordance with $|w_{kk}| < n^{1/4}$ or not, respectively. Then applying (2.26) and noticing $k_n = o(n)$, one gets the following contradiction:

$$\begin{aligned} &\lambda_{\max}(n^{-1/2}\mathbf{W}) \ge n^{-1/2}\mathbf{x}^*\mathbf{W}\mathbf{x} \\ &\ge (n - k_n - 1)^{1/2}\mathrm{E}(\mathrm{Re}(w_{12})) - n^{-1/4} + \lambda_{\min}(n^{-1/2}[\widetilde{\mathbf{W}} - \mathrm{E}(\widetilde{\mathbf{W}})]) \to \infty, \end{aligned}$$

where $\widetilde{\mathbf{W}}$ is the matrix obtained from $\mathbf{W}$ with its diagonal elements replaced by zero. Here, we have used the fact that $\lambda_{\min}(n^{-1/2}[\widetilde{\mathbf{W}} - \mathrm{E}(\widetilde{\mathbf{W}})]) \to -2\sigma^2$, by the sufficiency part of the theorem. This proves that the real parts of the means of off-diagonal elements of $\mathbf{W}$ cannot be positive.

If $b = \mathrm{E}(\mathrm{Im}(w_{12})) \ne 0$, define $\mathbf{x}$ in such a way that $x_j = 0$ if $|w_{jj}| > n^{1/4}$, and the other $n - k_n$ elements are $(n - k_n)^{-1/2}(1, e^{i\pi \mathrm{sign}(b)(2\ell-1)/(n-k_n)}, \ldots, e^{i\pi \mathrm{sign}(b)(2\ell-1)(n-k_n-1)/(n-k_n)})$, respectively. Note that $\mathbf{x}$ is the eigenvector corresponding to the eigenvalue $\cot(\pi(2\ell - 1)/2(n - k_n))$ of the Hermitian matrix whose (j, k)th $(j < k)$ element is i if $|w_{jj}| \le n^{1/4}$ and $|w_{kk}| \le n^{1/4}$, or 0 otherwise. Therefore, we have, with $a = |\mathrm{E}(\mathrm{Re}(w_{12})|$,

$$\begin{aligned} &\lambda_{\max}(n^{-1/2}\mathbf{W}) \ge n^{-1/2}\mathbf{x}^*\mathbf{W}\mathbf{x} \\ &\ge -\frac{|a|}{(n - k_n)\sqrt{n}\sin^2(\pi(2\ell - 1)/2(n - k_n))} + \frac{|b|}{\sqrt{n}\sin(\pi(2\ell - 1)/2(n - k_n))} \\ &\quad + \lambda_{\min}\Big(n^{-1/2}(\widetilde{\mathbf{W}} - \mathrm{E}(\widetilde{\mathbf{W}}))\Big) - n^{-1/4} \\ &:= I_1 + I_2 + I_3 - n^{-1/4}. \end{aligned}$$

Taking $\ell = [n^{1/3}]$ and noticing $k_n = o(n)$, we have

$$I_1 \sim -|a|n^{-1/6} \to 0, \quad I_2 \sim |b|n^{1/6} \to \infty \quad \text{and} \quad I_3 \to -2\sigma^2.$$

This leads to the contradiction that $\lambda_{\max}(n^{-1/2}\mathbf{W}) \to \infty$, proving the necessity of Condition (ii).

Condition (iii) follows by applying the sufficiency part. The proof of Theorem 2.12 is now complete.

Remark 2.6. For the Wigner matrix, there is a symmetry between the largest and smallest eigenvalues. Thus, Theorem 2.12 actually proves that the necessary and sufficient conditions for both the largest and smallest eigenvalues to have finite limits almost surely are that the diagonal elements have finite second moments and the off-diagonal elements have zero mean and finite fourth moments.

Remark 2.7. In the proof of Theorem 2.12, if the entries of $\mathbf{W}$ depend on n but satisfy

$$\mathrm{E}(w_{jk}) = 0, \ \mathrm{E}(|w_{jk}^2|) \leq \sigma^2, \ \mathrm{E}(|w_{jk}^\ell|) \leq b(\delta_n n)^{\ell-3}, \ (\ell \geq 3) \tag{2.31}$$

for some $b > 0$ and $\delta_n \downarrow 0$, then for fixed $\varepsilon > 0$ and $\ell > 0$, the following is true:

$$\mathrm{P}(\lambda_{\max}(n^{-1/2}\mathbf{W}) \geq 2\sigma + \varepsilon + x) = o(n^{-\ell}(2\sigma + \varepsilon + x)^{-2}), \tag{2.32}$$

uniformly for $x > 0$. This implies that the conclusion $\limsup \lambda_{\max}(n^{-1/2}\mathbf{W}) \leq 2\sigma$ a.s. is still true.

2.2.2. Limits of extreme eigenvalues of sample covariance matrices

Geman (1980) proved that, as $p/n \to y$, the largest eigenvalue of a sample covariance matrix tends to $b(y)$ almost surely, assuming a certain growth condition on the moments of the underlying distribution, where $b(y) = \sigma^2(1+\sqrt{y})^2$ is defined in the statement of Theorem 2.5. Later, Yin, Bai and Krishnaiah (1988) and Bai, Silverstein and Yin (1988), respectively, proved that the necessary and sufficient condition for the largest eigenvalue of a sample covariance matrix to converge to a finite limit almost surely is that the underlying distribution has a zero mean and finite fourth moment, and that the limit must be $b(y)$. Silverstein (1989b) showed that the necessary and sufficient conditions for the weak convergence of the largest eigenvalue of a sample covariance matrix are $E(x_{11}) = 0$ and $n^2\mathrm{P}(|x_{11}| \geq \sqrt{n}) \to 0$. The most difficult problem in this direction is to establish the strong convergence of the smallest eigenvalue of a sample covariance matrix. Yin, Bai and Krishnaiah (1983) and Silverstein (1984) showed that when $y \in (0,1)$, there is a positive constant ε_0 such that the liminf of the smallest eigenvalue of $1/n$ times a Wishart matrix is larger than ε_0, a.s. In Silverstein (1985), this result is further improved to say that the smallest eigenvalue of a normalized Wishart matrix tends to $a(y) = \sigma^2(1-\sqrt{y})^2$ almost surely. Silverstein's approach strongly relies on the normality assumption and hence cannot be extended to the general case. The latest contribution is due to Bai and Yin (1993), in which a unified approach is presented, establishing the strong convergence of both the largest and smallest eigenvalues simultaneously under the existence of the fourth moment. Although only the real case is considered in Bai and Yin (1993), their results can easily be extended to the complex case.

Theorem 2.15. *In addition to the assumptions of Theorem* 2.5, *we assume that the entries of* $\mathbf{X}$ *have finite fourth moment. Then*

$$-2y\sigma^2 \le \liminf_{n\to\infty} \lambda_{\min}(\mathbf{S}-\sigma^2(1+y)\mathbf{I}) \le \limsup_{n\to\infty} \lambda_{\max}(\mathbf{S}-\sigma^2(1+y)\mathbf{I}) \le 2y\sigma^2, \quad \text{a.s.} \tag{2.33}$$

If we define the smallest eigenvalues as the $(p-n+1)$-st smallest eigenvalue of $\mathbf{S}$ when $p > n$, then from Theorem 2.15, one immediately gets the following Theorem.

Theorem 2.16. *Under the assumptions of Theorem* 2.15, *we have*

$$\lim_{n\to\infty} \lambda_{\min}(\mathbf{S}) = \sigma^2(1-\sqrt{y})^2, \quad a.s. \tag{2.34}$$

and

$$\lim_{n\to\infty} \lambda_{\max}(\mathbf{S}) = \sigma^2(1+\sqrt{y})^2, \quad \text{a.s.} \tag{2.35}$$

The proof of Theorem 2.15 relies on the following two lemmas.

Lemma 2.17. *Under the conditions of Theorem* 2.15, *we have*

$$\limsup_{n\to\infty} ||\mathbf{T}(\ell)|| \le (2\ell+1)(\ell+1)y^{(\ell-1)/2} \quad a.s.,$$

where $\mathbf{T}(\ell)$, $p\times p$, *has its* (a,b)*th entry* $n^{-\ell}(\sum' x_{av_1}\overline{x}_{u_1v_1}x_{u_1v_2}\overline{x}_{u_2v_2}\cdots x_{u_{\ell-1}v_\ell}\overline{x}_{bv_\ell})$ *and the summation* $\sum'$ *runs over* $v_1,\dots,v_\ell = 1,\dots,n$ *and* $u_1,\dots,u_{\ell-1} = 1,\dots,p$ *subject to the restriction*

$$a \ne u_1,\ u_1 \ne u_2,\dots,u_{\ell-1} \ne b \quad and \quad v_1 \ne v_2,\ v_2 \ne v_3,\dots,v_{\ell-1} \ne v_\ell.$$

Lemma 2.18. *Under the conditions of Theorem* 2.15, *we have*

$$(\mathbf{T}-y\mathbf{I})^h = \sum_{r=0}^{h}(-1)^{r+1}\mathbf{T}(r)\sum_{i=0}^{[(h-r)/2]} C_i(h,r)y^{h-r-i} + o(1), \tag{2.36}$$

where $\mathbf{T} = \mathbf{T}(1) = \mathbf{S}-\sigma^2(1+y)\mathbf{I}$ *and the constants* $|C_i(h,r)| \le 2^h$.

The proof of Lemma 2.17 is similar to that of Theorem 2.12, i.e., to consider the expectation of $\text{tr}(\mathbf{T}^{2s}(\ell))$. Construct the graphs as in the proof of Theorem 2.5. Using Lemmas 2.13 and 2.14 one gets an estimate

$$\mathrm{E}(\text{tr}(\mathbf{T}^{2s}(\ell))) \le n^3[(2\ell+1)(\ell+1)y^{(\ell-1)/2} + o(1)]^{2s}.$$

From this, Lemma 2.17 can be proved; the details are omitted. The proof of Lemma 2.18 follows by induction.

2.3. Limiting behavior of eigenvectors

Relatively less work has been done on the limiting behavior of eigenvectors than eigenvalues in the spectral analysis of LDRM. Some work on eigenvectors of the Wigner matrix can be found in Girko, Kirsch and Kutzelnigg (1994), in which the first order properties are investigated. For eigenvectors of sample covariance matrices, some results can be found in Silverstein (1979, 1981, 1984b, 1989, 1990). Except for his first paper, the focus is on second order properties.

There is a good deal of evidence that the behavior of LDRM is asymptotically distribution-free, that is, it is asymptotically equivalent to the case where the basic entries are i.i.d. mean 0 normal, provided certain moment requirements are met. This phenomenon has been confirmed for distributions of eigenvalues. For the eigenvectors, the problem is how to formulate such a property. In the normal case, the matrix of orthonormal eigenvectors, which will be simply called the eigenmatrix, is Haar distributed. Since the dimension tends to infinity, it is difficult to compare the distribution of the eigenmatrix with the Haar measure. However, there are several different ways to characterize the similarity between these two distributions. The following approach is considered in the work referred to above.

Let $\mathbf{u}_n = (u_1, \ldots, u_p)'$ be a p-dimensional unit vector and $\mathbf{O}_n$ be the eigenmatrix of a covariance matrix. Define $\mathbf{y}_n = \mathbf{O}_n'\mathbf{u}_n = (y_1, \ldots, y_p)'$. If $\mathbf{O}_n$ is Haar distributed, then $\mathbf{y}$ is uniformly distributed on the unit sphere in a p-dimensional space. To this end, define a stochastic process $Y_n(t)$ as follows.

$$Y_n(t) = \sum_{i=1}^{[pt]} |y_i|^2.$$

Note that the process can also be viewed as a random measure of the uniformity of the distribution of $\mathbf{y}$. It is conceivable that $Y_n(F_n(t))$ converges to a common limiting stochastic process whatever the vector $\mathbf{u}_n$ is, where F_n is the ESD of the random matrix. This was proved in Girko, Kirsch and Kutzelnigg (1994) for the Wigner matrix and was the the main focus of Silverstein (1979) for large covariance matrices. This is implied by results in Silverstein's other work, in which second order properties are investigated. Here, we shall briefly introduce some of his results in this direction.

In the remainder of this subsection, we consider a real sample covariance matrix $\mathbf{S}$ with i.i.d. entries. Define

$$X_n(t) = \sqrt{p/2}(Y_n(t) - [pt]/p).$$

When $\mathbf{S}$ is a Wishart matrix, it is not difficult to show that $X_n(t)$ converges weakly to a Brownian bridge $W^0(t)$ in $D[0,1]$, the space of r.c.l.l. (right-continuous and left-limit) functions on $[0,1]$. In Silverstein (1989a), the following theorem is proved.

Theorem 2.19.

(i) *If*

$$\mathrm{E}(x_{11}) = 0,\ \mathrm{E}(|x_{11}^2|) = 1,\quad \mathrm{E}(|x_{11}^4|) = 3, \tag{2.37}$$

then for any integer k

$$\Big(\int_0^\infty x^r\, X_n(F^{\mathbf{S}}(dx)),\ r = 1,\ldots,k\Big) \xrightarrow{\mathcal{D}} \Big(\int_0^\infty x^r\, W^0(F_y(dx)),\ r = 1,\ldots,k\Big), \tag{2.38}$$

where F_y is the Marčenko-Pastur distribution with dimension-ratio y and parameter $\sigma^2 = 1$.

(ii) *If* $\int_0^\infty x\, X_n(F^{\mathbf{S}}(dx))$ *is to converge in distribution to a random variable for* $\mathbf{u}_n = (1,0,0,\ldots,0)'$ *and* $\mathbf{u}_n = p^{-1/2}(1,1,\ldots,1)'$, *then* $\mathrm{E}(|x_{11}^4|) < \infty$ *and* $\mathrm{E}(x_{11}) = 0$.

(iii) *If* $\mathrm{E}(|x_{11}^4|) < \infty$ *but* $\mathrm{E}(|x_{11} - \mathrm{E}(x_{11})|^4)/\mathrm{Var}(x_{11}) \neq 3$, *then there exist sequences* $\{\mathbf{u}_n\}$ *for which*

$$\Big(\int_0^\infty x\, X_n(F^{\mathbf{S}}(dx)),\ \int_0^\infty x^2\, X_n(F^{\mathbf{S}}(dx))\Big)$$

fails to converge in distribution.

Note that

$$\int_0^\infty x^k\, X_n(F^{\mathbf{S}}(dx)) = \sqrt{p/2}\Big(\mathbf{u}_n^*\mathbf{S}^k\mathbf{u}_n - p^{-1}\mathrm{tr}(\mathbf{S}^k)\Big).$$

The proof of (i) consists of the following three steps

1) $\sqrt{p}\mathrm{E}(\mathbf{u}_n^*\mathbf{S}^k\mathbf{u}_n - p^{-1}\mathrm{tr}(\mathbf{S}^k)) \xrightarrow{Pr} 0$;

2) $\sqrt{p}(p^{-1}\mathrm{tr}(\mathbf{S}^k) - \mathrm{E}(p^{-1}\mathrm{tr}(\mathbf{S}^k))) \xrightarrow{Pr} 0$;

3) $\sqrt{p/2}(\mathbf{u}_n^*\mathbf{S}^k\mathbf{u}_n - \mathrm{E}(\mathbf{u}_n^*\mathbf{S}^k\mathbf{u}_n)) \xrightarrow{\mathcal{D}} \int_0^\infty x^k\, W^0(F_y(dx))$.

The details are omitted. The proof of (ii) follows from standard limit theorems (see, e.g., Gnedenko and Kolmogorov (1954)). As for conclusion (iii), by elementary computation we have

$$\begin{aligned}&\mathrm{Cov}\Big(\int_0^\infty x\, X_n(F^{\mathbf{S}}(dx)),\ \int_0^\infty x^2\, X_n(F^{\mathbf{S}}(dx))\Big)\\ &= (2y + y^2)(1 + o(1))\Big[1 + \frac{1}{2}(\mathrm{E}(|x_{11}^4|) - 3)\Big(\sum_{i=1}^n |\mathbf{u}_i|^4\Big)\Big].\end{aligned}$$

Then $\mathbf{u}_n$ can be chosen so that the right hand side of the above has no limit, unless $\mathrm{E}(|x_{11}^4|) = 3$.

Remark 2.8. The importance of the above theorem stems from the following. Assume $\mathrm{Var}\,(x_{11}) = 1$. If $\mathrm{E}(x_{11}) = 0$, $n^2\mathrm{P}(|x_{11}| \geq \sqrt{n}) \to 0$ (ensuring weak convergence of the largest eigenvalue of $\mathbf{S}$) and $X_n \xrightarrow{\mathcal{D}} W^0$, then it can be shown that (2.38) holds. Therefore, if weak convergence to a Brownian bridge is to hold for all choices of unit vectors $\mathbf{u}$, from (ii) and (iii) it must follow that $\mathrm{E}(|x_{11}^4|) = 3$. Thus it appears that similarity of the eigenmatrix to the Haar measure requires a certain amount of closeness of x_{11} to the standard normal distribution. At present, either of the two extremes, $X_n \xrightarrow{\mathcal{D}} W^0$ for all unit $\mathbf{u}$ and all x_{11} satisfying the above moment conditions, or $X_n \xrightarrow{\mathcal{D}} W^0$ only in the Wishart case, remains as a possibility. However, because of (i), verifying weak convergence to a Brownian bridge amounts to showing tightness of the sequence $\{X_n\}$ in $D[0, 1]$.

The following theorem, found in Silverstein (1990), yields a partial solution to the problem, a case where tightness can be established.

Theorem 2.20. *Assume x_{11} is symmetrically distributed about* 0 *and* $\mathrm{E}(x_{11}^4) < \infty$. *Then $X_n \xrightarrow{\mathcal{D}} W^0$ holds for $\mathbf{u} = p^{-1/2}(\pm 1, \pm 1, \ldots)$.*

2.4. Miscellanea

Let $\mathbf{X}$ be an $n \times n$ matrix of i.i.d. complex random variables with mean zero and variance σ^2. In Bai and Yin (1986), large systems of linear equations and linear differential equations are considered. There, the norm of $(n^{-1/2}\mathbf{X})^k$ plays an important role for the stability of the solutions. The following theorem was proved.

Theorem 2.21. *If* $\mathrm{E}(|x_{11}^4|) < \infty$, *then*

$$\limsup_{n\to\infty} ||(n^{-1/2}\mathbf{X})^k|| \leq (1+k)\sigma^k, \quad \text{a.s., for all } k. \tag{2.39}$$

The proof of this theorem relies on, after truncation and centralization, the estimation of $\mathrm{E}([(n^{-1/2}\mathbf{X})^k(n^{-1/2}\mathbf{X}^*)^k]^\ell)$. The details are omitted. Here, we remark that when $k = 1$, the theorem reduces to a special case of Theorem 2.15 for $y = 1$. We also introduce an important consequence about the spectral radius of $n^{-1/2}\mathbf{X}$, which plays an important role in establishing the circular law (See Section 4). This was also independently proved by Geman (1986), under additional restrictions on the growth of moments of the underlying distribution.

Theorem 2.22. *If* $\mathrm{E}(|x_{11}^4|) < \infty$, *then*

$$\limsup_{n\to\infty} \max_{j\leq n} |\lambda_j(n^{-1/2}\mathbf{X})| \leq \sigma, \quad \text{a.s.} \tag{2.40}$$

Theorem 2.22 follows from the fact that for any k,

$$\limsup_{n\to\infty}\max_{j\le n}|\lambda_j(n^{-1/2}\mathbf{X})| = \limsup_{n\to\infty}\max_{j\le n}|\lambda_j[(n^{-1/2}\mathbf{X})^k]|^{1/k}$$
$$\le \limsup_{n\to\infty}||(n^{-1/2}\mathbf{X})^k||^{1/k} \le (1+k)^{1/k}\sigma \to \sigma,$$

by making $k\to\infty$.

Remark 2.9. Checking the proof of Theorem 2.21, one finds that, after truncation and centralization, the conditions for guaranteeing (2.39) are $|x_{jk}|\le\delta_n\sqrt{n}$, $\mathrm{E}(|x_{jk}^2|)\le\sigma^2$ and $\mathrm{E}(|x_{jk}^3|)\le b$, for some $b>0$. This is useful in extending the circular law to the case where the entries are not identically distributed.

3. Stieltjes Transform

Let G be a function of bounded variation defined on the real line. Then its Stieltjes transform is defined by

$$m(z) = \int_{-\infty}^{\infty}\frac{1}{x-z}G(dx), \tag{3.1}$$

where $z=u+iv$ with $v>0$. Throughout this section, z denotes $u+iv$ with $v>0$. Note that the integrand in (3.1) is bounded by $1/v$, the integral always exists, and

$$\frac{1}{\pi}\mathrm{Im}(m(z)) = \int_{-\infty}^{\infty}\Big[\frac{v}{\pi[(x-u)^2+v^2]}\Big]G(dx).$$

This is the convolution of G with a Cauchy density with a scale parameter v. If G is a distribution function, then the Stieltjes transform always has a positive imaginary part. Thus, one can easily verify that, for any continuity points $x_1 < x_2$ of G,

$$\lim_{v\downarrow 0}\int_{x_1}^{x_2}\frac{1}{\pi}\mathrm{Im}(m(z))du = G(x_2)-G(x_1). \tag{3.2}$$

Formula (3.2) obviously provides a continuity theorem between the family of distribution functions and the family of their Stieltjes transforms.

Also, if $\mathrm{Im}(m(z))$ is continuous at x_0+i0, then $G(x)$ is differentiable at $x=x_0$ and its derivative equals $\frac{1}{\pi}\mathrm{Im}(m(x_0+i0))$. This result was stated in Bai (1993a) and rigorously proved in Silverstein and Choi (1995). Formula (3.2) gives an easy way to find the density of a distribution function if its Stieltjes transform is known.

Now, let G be the ESD of a Hermitian matrix $\mathbf{W}$ of order p. Then it is easy to see that

$$m_G(z) = \frac{1}{p}\mathrm{tr}(\mathbf{W}-z\mathbf{I})^{-1}$$
$$= \frac{1}{p}\sum_{k=1}^{p}\frac{1}{w_{kk}-z-\boldsymbol{\alpha}_k^*(\mathbf{W}_k-z\mathbf{I}_{p-1})^{-1}\boldsymbol{\alpha}_k}, \tag{3.3}$$

where $\boldsymbol{\alpha}_k$ $((p-1)\times 1)$ is the kth column vector of $\mathbf{W}$ with the kth element removed and $\mathbf{W}_k$ is the matrix obtained from $\mathbf{W}$ with the kth row and column deleted. Formula (3.3) provides a powerful tool in the area of spectral analysis of LDRM.

As mentioned earlier, the mapping from distribution functions to their Stieltjes transforms is continuous. In Bai (1993a), this relation was more clearly characterized as a Berry-Esseen type inequality.

Theorem 3.1. *Let F be a distribution function and G be a function of bounded variation satisfying $\int |F(y)-G(y)|dy<\infty$. Then, for any $v>0$ and constants γ and a related to each other by the condition $\gamma=\frac{1}{\pi}\int_{|u|\le a}\frac{1}{u^2+1}du>1/2$,*

$$||F-G||\le\frac{1}{\pi(2\gamma-1)}\Big[\int|f(z)-g(z)|du+\frac{1}{v}\sup_x\int_{|y|\le 2va}|G(x+y)-G(x)|dy\Big],$$

where f and g are Stieltjes transforms of F and G respectively, and $z=u+iv$.

Sometimes, F and G have thin tails or even have bounded supports. In these cases, one may want to bound the difference between F and G in terms of an estimate of the difference of their Stieltjes transforms on a finite interval. We have the following theorem.

Theorem 3.2. *Under the conditions of Theorem 3.1, for any constants A and B restricted by $\kappa=\frac{4B}{\pi(A-B)(2\gamma-1)}\in(0,1)$, we have*

$$\|F-G\|\le\frac{1}{\pi(1-\kappa)(2\gamma-1)}\Big[\int_{-A}^{A}|f(z)-g(z)|du$$
$$+\frac{1}{v}\sup_x\int_{|y|\le 2va}|G(x+y)-G(x)|dy+2\pi v^{-1}\int_{|x|>B}|F(x)-G(x)|dx\Big].$$

Corollary 3.3. *In addition to the conditions of Theorem 3.1, assume further that, for some constant B, $F([-B,B])=1$ and $|G|((-\infty,-B))=|G|((B,\infty))=0$, where $|G|((b,c))$ denotes the total variation of G on the interval (b,c). Then for any A satisfying the constraint in Thereom 3.2, we have*

$$\|F-G\|\le\frac{1}{\pi(1-\kappa)(2\gamma-1)}\Big[\int_{-A}^{A}|f(z)-g(z)|du+\frac{1}{v}\sup_x\int_{|y|\le 2va}|G(x+y)-G(x)|dy\Big].$$

Remark 3.1. Corollary 3.3 is good enough for establishing the convergence rate of ESD's of LDRM since, in all known cases in the literature, the limiting distribution has a bounded support and the extreme eigenvalues have finite limits. It is more convenient than Theorem 3.1 since one does not need to estimate the integral of the difference of the Stieltjes transforms over the whole line.

3.1. Limiting spectral distributions

As an illustration, we use the Stieltjes transform (3.3) to derive the LSD's of Wigner and sample covariance matrices.

3.1.1. Wigner matrix

Now, as an illustration of how to use Formula (3.3) to find the LSD's, let us give a sketch of the proof of Theorem 2.1. Truncation and centralization are done first as in the proof of Theorem 2.1. That is, we may assume that $w_{kk} = 0$ and $|w_{jk}| \leq C$ for all $j \neq k$ and some constant C. Theorem 2.4 can be similarly proved but needs more tedious arguments.

Let $m_n(z)$ be the Stieltjes transform of the ESD of $n^{-1/2}\mathbf{W}$. By (3.3), and noticing $w_{kk} = 0$, we have

$$\begin{aligned} m_n(z) &= \frac{1}{n}\sum_{k=1}^{n} \frac{1}{-z - \frac{1}{n}\boldsymbol{\alpha}_k^*(n^{-1/2}\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k} \\ &= \frac{1}{n}\sum_{k=1}^{n} \frac{1}{-z - \sigma^2 m_n(z) + \varepsilon_k} = -\frac{1}{z + \sigma^2 m_n(z)} + \delta_n, \end{aligned} \tag{3.4}$$

where

$$\varepsilon_k = \sigma^2 m_n(z) - \frac{1}{n}\boldsymbol{\alpha}_k^*(n^{-1/2}\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k$$

and

$$\delta_n = \delta_n(z) = \frac{1}{n}\sum_{k=1}^{n} \frac{-\varepsilon_k}{(-z - \sigma^2 m_n(z) + \varepsilon_k)(-z - \sigma^2 m_n(z))}. \tag{3.5}$$

We first show that for any fixed $v_0 > 0$ and $B > 0$, with $z = u + iv$,

$$\sup_{|u|\leq B, v_0 \leq v \leq B} |\delta_n(z)| = o(1) \text{ a.s.} \tag{3.6}$$

By the uniform continuity of $m_n(z)$, the proof of (3.6) is equivalent to showing for each fixed z with $v > 0$,

$$|\delta_n(z)| = o(1) \text{ a.s.} \tag{3.7}$$

Note that

$$\begin{aligned} &| - z - \sigma^2 m_n(z) + \varepsilon_k| \geq \mathrm{Im}(-z - \frac{1}{n}\boldsymbol{\alpha}_k^*(n^{-1/2}\mathbf{W}_k - z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k)| \\ &= v(1 + \frac{1}{n}\boldsymbol{\alpha}_k^*((n^{-1/2}\mathbf{W}_k - u\mathbf{I}_{n-1})^2 + v^2\mathbf{I})^{-1}\boldsymbol{\alpha}_k) \geq v, \end{aligned}$$

and $|z + \sigma^2 m_n(z)| \geq v$. Then (3.7) follows if one can show

$$\max_k |\varepsilon_k(z)| = o(1) \text{ a.s.} \tag{3.8}$$

Let F_n and $F_{n(-k)}$ denote the ESD's of $n^{-1/2}\mathbf{W}$ and $n^{-1/2}\mathbf{W}_k$, respectively. Since $|nF_n(x)-(n-1)F_{n(-k)}(x)|\le 1$ by the interlacing theorem (see the proof of Lemma 2.2),

$$|m_n(z)-\frac{1}{n}\mathrm{tr}((n^{-1/2}\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1})|=\Big|\frac{1}{n}\int\frac{nF_n(dx)-(n-1)F_{n(-k)}(dx)}{x-z}\Big|$$
$$\le\Big|\frac{1}{n}\int\frac{(nF_n(x)-(n-1)F_{n(-k)}(x))dx}{(x-z)^2}\Big|\le\pi/nv.$$

Based on this fact, in the proof of (3.8), we can replace ε_k's by $n^{-1}\boldsymbol{\alpha}_k^*(n^{-1/2}\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k)-n^{-1}\mathrm{tr}((n^{-1/2}\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1})$.

Since $\boldsymbol{\alpha}_k$ is independent of $\mathbf{W}_k$, it is not difficult to show that

$$\mathrm{E}\Big(\Big|n^{-1}\boldsymbol{\alpha}_k^*(n^{-1/2}\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1}\boldsymbol{\alpha}_k-n^{-1}\mathrm{tr}(n^{-1/2}\mathbf{W}_k-z\mathbf{I}_{n-1})^{-1})\Big|^6\Big|\mathbf{W}_k\Big)$$
$$\le\frac{6!C^6}{n^6}\Big(\mathrm{tr}\Big[(n^{-1/2}\mathbf{W}_k-u\mathbf{I}_{n-1})^2+v^2\mathbf{I}_{n-1}\Big]^{-1})\Big)^3=O(n^{-3}).$$

This implies (3.8).

Solving equation (3.4) (in the variable m), one gets two solutions

$$m_n^{(1),(2)}(z)=-\frac{1}{2\sigma^2}[z+\delta_n\sigma^2\mp\sqrt{(z-\delta_n\sigma^2)^2-4\sigma^2}],$$

where, for a complex number a, by convention $\sqrt{a}$ denotes the square root with positive imaginary part. We need to determine which solution is the Stieltjes transform of the spectrum of $n^{-1/2}\mathbf{W}$. By (3.4), we have

$$|\delta_n|\le|m_n|+1/|z+\sigma^2m_n|\le 2/v\to 0,\quad\text{as}\quad v\to\infty.$$

Thus, when z has a large imaginary part, $m_n=m_n^{(1)}(z)$. We claim this is true for all z with $v>0$. Note that m_n and $m_n^{(1),(2)}$ are continuous in z on the upper half complex plane. We only need to show that $m_n^{(1)}$ and $m_n^{(2)}$ have no intersection. Suppose that they are equal at z_0 with $\mathrm{Im}(z_0)>0$. Then we have $(z_0-\sigma^2\delta_n)^2-4\sigma^2=0$ and

$$m_n(z_0)=-\frac{1}{2\sigma^2}(z_0+\sigma^2\delta_n)=-z_0/\sigma^2\pm 2/\sigma,$$

which contradicts with the fact that $m_n(z)$ has a positive imaginary part. Therefore, we have proved that

$$m_n(z)=-\frac{1}{2\sigma^2}[z+\delta_n\sigma^2-\sqrt{(z-\delta_n\sigma^2)^2-4\sigma^2}].$$

Then from (3.6), it follows that with probability 1 for every fixed z with $v>0$, $m_n(z)\to m(z)=-\frac{1}{2\sigma^2}[z-\sqrt{z^2-4\sigma^2}]$. Letting $v\downarrow 0$, we find the density of semicircular law as give in (2.2).

3.1.2. General sample covariance matrix

Note that a general form of sample covariance matrices can be considered as a special case of products of random matrices $\mathbf{ST}$ in Theorem 2.10. For generalization in another direction, as mentioned in Section 2.1.3, we present the following theorem.

Theorem 3.4. (Silverstein and Bai (1995)) *Suppose that for each* n*, the entries of* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$*,* $p \times n$*, are i.i.d. complex random variables with* $\mathrm{E}(|x_{11} - \mathrm{E}(x_{11})|^2) = 1$*, and that* $\mathbf{T} = \mathbf{T}_n = \mathrm{diag}(\tau_1^n, \ldots, \tau_p^n)$*,* τ_i^n *real, and the ESD of* $\mathbf{T}$ *converges almost surely to a probability distribution function* H *as* $n \to \infty$*. Assume that* $\mathbf{B} = \mathbf{A} + \frac{1}{n}\mathbf{X}^*\mathbf{T}\mathbf{X}$*, where* $\mathbf{A} = \mathbf{A}_n$ *is Hermitian* $n \times n$ *satisfying* $F^{\mathbf{A}_n} \xrightarrow{v} F_a$ *almost surely, where* F_a *is a distribution function (possibly defective, i.e., of total variation less than* 1*) on the real line, and* $\xrightarrow{v}$ *means vague convergence, i.e., convergence without preservation of the total variation. Furthermore, assume that* $\mathbf{X}$*,* $\mathbf{T}$*, and* $\mathbf{A}$ *are independent. When* $p/n \to y > 0$ *as* $n \to \infty$*, we have almost surely* $F^{\mathbf{B}}$*, the ESD of* $\mathbf{B}$*, converges vaguely to a (non-random) d.f.* F*, whose Stieltjes transform* $m(z)$ *is given by*

$$m(z) = m_a\Big(z - y\int \frac{\tau dH(\tau)}{1 + \tau m(z)}\Big), \tag{3.9}$$

where z *is a complex number with a positive imaginary part and* m_a *is the Stieltjes transform of* F_a*.*

The set-up of Theorem 3.4 originated from nuclear physics, but is also encountered in multivariate statistics. In MANOVA, $\mathbf{A}$ can be considered as the between-covariance matrix, which may diverge in some directions under the alternative hypothesis. Examples of $\mathbf{B}$ can be found in the analysis of multivariate linear models and error-in-variables models, when the sample covariance matrix of the covariates is ill-conditioned. The role of $\mathbf{A}$ is to reduce the instability in the directions of the eigenvectors corresponding to small eigenvalues.

Remark 3.2. Note that Silverstein and Bai (1995) is more general than Yin (1986) in that it does not require the moment convergence of the ESD of $\mathbf{T}$ nor the positive definiteness of $\mathbf{T}$. Also, it allows a perturbation matrix $\mathbf{A}$. However, it is more restrictive than Yin (1986) in that it requires the matrix $\mathbf{T}$ to be diagonal. An extension of Yin's work in another direction is made in Silverstein (1995), who only assumes that $\mathbf{T}$ is positive definite and its ESD almost surely tends to a probability distribution, without requiring moment convergence. Weak convergence to (3.9) was established in Marčenko-Pastur (1967) under higher moment conditions than assumed in Theorem 3.4, but with mild dependence between the entries of $\mathbf{X}$.

The assumption that the matrix $\mathbf{T}$ is diagonal in Theorem 3.4 is needed for the proof. It seems possible and is of interest to remove this restriction.

Now, we sketch a proof of Theorem 3.4 under more general conditions by using the Stieltjes transform. We replace the conditions for the x-variables with those given in Theorem 2.8. Remember that the entries of $\mathbf{X}$ and $\mathbf{T}$ depend on n. For brevity, we shall suppress the index n from these symbols and τ_i^n.

Denote by H_n and H the ESD of $\mathbf{T}_n$ and its LSD, and denote by m_{A_n} and m_A the Stieltjes transforms of the ESD of $\mathbf{A}_n$ and that of its LSD. Denote the Stieltjes transform of the ESD of $\mathbf{B}$ by $m_n(z)$.

Using the truncation and centralization techniques as in the proof of Theorem 2.10, without loss of generality, we may assume that the following additional conditions hold:

1. $|\tau_j| \le \tau_0$ for some positive constant τ_0,
2. $\mathrm{E}(x_{ij}) = 0$, $\mathrm{E}(|x_{ij}|^2) \le 1$ with $\frac{1}{pn}\sum_{ij}\mathrm{E}(|x_{ij}|^2) \to 1$ and $|x_{ij}| \le \delta_n\sqrt{n}$ for some sequence $\delta_n \to 0$.

If $F^{\mathbf{A}_n} \to c$, a.s. for some $c \in [0,1]$ (which is equivalent to almost all eigenvalues of $\mathbf{A}_n$ tending to infinity while the number of eigenvalues tending to negative infinity is about cn), then $F^{\mathbf{B}_n} \to c$ a.s., since the support of $\mathbf{XTX}^*$ remains bounded. Consequently, $m_n \to 0$ and $m_{A_n} \to 0$ a.s., and hence (3.9) is true. Thus, we only need to consider the case where the limit F^A of F^{A_n} has a positive mass over the real line. Then for any fixed z, there is a positive number η such that $\mathrm{Im}(m_n(z)) > \eta$.

Let $\mathbf{B}_{(i)} = \mathbf{B} - \tau_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^*$ and

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i}{1+\tau_i m_n(z)} = \int \frac{\tau}{1+\tau m_n(z)}H_n(d\tau),$$

where $\boldsymbol{\xi}_i = n^{-1/2}\mathbf{x}_i$. Note that x has a non-positive imaginary part. Then by the identity

$$(\mathbf{A}_n-(z-\mu_n)\mathbf{I})^{-1} = (\mathbf{B}-z\mathbf{I})^{-1} + (\mathbf{A}_n-(z-\mu_n)\mathbf{I})^{-1}(\frac{1}{n}\mathbf{XTX}^* - \mu_n\mathbf{I})(\mathbf{B}-z\mathbf{I})^{-1},$$

we obtain

$$m_{A_n}(z-\mu_n) - m_n(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i d_i}{1+\tau_i m_n(z)}, \tag{3.10}$$

where

$$\begin{aligned} d_i = d_i^n(\mu_n) = {} & \frac{(1+\tau_i m_n(z))\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}(\mathbf{A}_n-(z-\mu_n)\mathbf{I})^{-1}\boldsymbol{\xi}_i}{1+\tau_i\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}\boldsymbol{\xi}_i} \\ & -\frac{1}{n}\mathrm{tr}[(\mathbf{B}-z\mathbf{I})^{-1}(\mathbf{A}_n-(z-\mu_n)\mathbf{I})^{-1}]. \end{aligned}$$

Note that for any fixed z, $\{m_n(z)\}$ is a bounded sequence. Thus, any subsequence of $\{m_n(z)\}$ contains a convergent subsequence. If m_n converges, then so does μ_n and hence $m_{A_n}(z-\mu_n)$. By (3.10), to prove (3.9), one only needs to show that equation (3.10) tends to (3.9) once $m_n(z)$ converges and that equation (3.9) has a unique solution. The proof of the latter is postponed to the next theorem. A proof of the former, i.e., the right hand side of (3.10) tends to zero, is presented here.

By (3.10) and the fact that $\mathrm{Im}(m_n(z)) > \eta$, we have $|1+\tau_i m_n(z)| \geq \min\{1/2, v\eta/2\tau_0\} > 0$. This implies that μ_n is uniformly bounded. Also, we know that μ_n has non-positive imaginary part from its definition. Therefore, to complete the proof of the convergence of (3.10), we can show the stronger conclusion that, with probability 1, the right hand side of (3.10) (with μ_n replaced by μ) tends to zero uniformly in μ over any compact set of the lower half complex plane. Due to the uniform continuity of both sides of (3.10) in u and μ, we only need to show (3.10) for any fixed z and non-random μ.

Note that the norms of $(A_n-(z-\mu)\mathbf{I})^{-1}$, $(\mathbf{B}-z\mathbf{I})^{-1}$ and $(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}$ are bounded by $1/v$. Now we present an easier proof under the slightly stronger condition that $\delta_n^2\log n\to 0$. (This holds if the random variables $|x_{jk}|^2\log(1+|x_{jk}|)$ are uniformly integrable or x_{jk} are identically distributed. For the second case, a second-step truncation is needed (see Silverstein and Bai (1995) for details)). Under this additional condition, it is sufficient to show that $\max_i\{|d_i|\}\to 0$, a.s. Using Lemma A.4 of Bai (1997), one can show that

$$\begin{aligned}&\mathrm{P}\Big(\Big|\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}(A_n-(z-\mu)\mathbf{I})^{-1}\boldsymbol{\xi}_i-\frac{1}{n}\mathrm{tr}[(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}(A_n-(z-\mu)\mathbf{I})^{-1}]\Big|\geq\varepsilon\Big)\\&\leq C\exp\{-b/\delta_n^2\}\end{aligned}$$

and

$$\mathrm{P}(|\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}\boldsymbol{\xi}_i-\frac{1}{n}\mathrm{tr}(\mathbf{B}_{(i)}-z\mathbf{I})^{-1})|\geq\varepsilon)\leq C\exp\{-b/\delta_n^2\},\quad\text{for some } b>0.$$

These two inequalities show that for any fixed μ,

$$\max_{i\leq p}|\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}(A_n-(z-\mu)\mathbf{I})^{-1}\boldsymbol{\xi}_i-n^{-1}\mathrm{tr}[(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}(A_n-(z-\mu)\mathbf{I})^{-1}]|\to 0,\quad\text{a.s.}$$

and

$$\max_{i\leq p}|\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}\boldsymbol{\xi}_i-\frac{1}{n}\mathrm{tr}(\mathbf{B}_{(i)}-z\mathbf{I})^{-1})|\to 0,\quad\text{a.s.}$$

Then the uniform continuity in μ implies that the above two limits hold uniformly when μ varies in any fixed compact subset of the lower half plane.

Note that the rank of $\mathbf{B}-\mathbf{B}_{(i)}$ is one. Also by Lemma 2.2, $||F^{\mathbf{B}}-F^{\mathbf{B}_{(i)}}||\leq 1/p$. Hence

$$|m_n(z)-m_{n(i)}(z)|\leq\frac{\pi}{pv}, \tag{3.11}$$

where $m_{n(i)}(z)$ is the Stieltjes transform of the ESD of $\mathbf{B}_{(i)}$. Therefore,

$$\max_{i\le p}\left|\frac{1+\tau_i m_n(z)}{1+\tau_i\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}\boldsymbol{\xi}_i}-1\right|=\max_{i\le p}\frac{o(1)}{1+\tau_i m_n(z)+o(1)}=o(1).$$

Here, the last step follows from the fact that $|1+\tau_i m_n(z)|\ge\min\{1/2,\, v\eta/2\tau_0\}>0$. Finally, we get

$$\begin{aligned}\max_{i\le p}|d_i|&\le o(1)+\max_{i\le p}\left|\frac{1}{n}\mathrm{tr}\left([(\mathbf{B}-z\mathbf{I})^{-1}-(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}](A_n-(z-\mu)\mathbf{I})^{-1}\right)\right|\\&\le o(1)+\max_{i\le p}\frac{1}{n^2v^3}\sum_{k=1}^n|x_{ik}|^2=o(1).\end{aligned}$$

For the general case without the additional condition, we need to rewrite the right hand side of (3.10) as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\frac{\tau_i d_{i1}}{1+\tau_i m_n(z)}-\frac{1}{n}\sum_{i=1}^n\frac{\tau_i^2\eta_i d_{i2}}{(1+\tau_i m_n(z))^2}-\frac{1}{n}\sum_{i=1}^n\frac{\tau_i^2 d_{i1}d_{i2}}{(1+\tau_i m_n(z))^2}\\&+\frac{1}{n}\sum_{i=1}^n\frac{\tau_i^3\eta_i d_{i2}^2}{(1+\tau_i m_n(z))^3}-\frac{1}{n}\sum_{i=1}^n\frac{\tau_i^4\eta_i d_{i2}^3}{(1+\tau_i m_n(z))^3(1+\tau_i\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}\boldsymbol{\xi}_i)}\\&+\frac{1}{n}\sum_{i=1}^n\frac{\tau_i^3 d_{i1}d_{i2}^2}{(1+\tau_i m_n(z))^2(1+\tau_i\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}\boldsymbol{\xi}_i)}+o_{a.s.}(n^{-1+\varepsilon})\\&=\sum_{k=1}^6\zeta_k+o_{a.s.}(n^{-1+\varepsilon}),\end{aligned}$$

where

$$d_{i1}=\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}(\mathbf{A}_n-(z-\mu)\mathbf{I})^{-1}\boldsymbol{\xi}_i-\eta_i,$$

$$d_{i2}=\boldsymbol{\xi}_i^*(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}\boldsymbol{\xi}_i-\frac{1}{n}\mathrm{tr}[(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}]$$

and

$$\eta_i=\frac{1}{n}\mathrm{tr}[(\mathbf{B}_{(i)}-z\mathbf{I})^{-1}(\mathbf{A}_n-(z-\mu)\mathbf{I})^{-1}].$$

By elementary but tedious arguments, one can show that $\mathrm{E}(|\zeta_k|^4)=O(n^{-2})$, $k=1,2$, $\mathrm{E}(|\zeta_k|^2)=O(n^{-2})$, $k=3,4$, and $\mathrm{E}(|\zeta_k|)=O(n^{-3/2})$, $k=5,6$. Thus, the right hand side of (3.10) tends to zero almost surely. The proof of Theorem 3.4 is complete.

Theorem 3.5. *For any z with* $\mathrm{Im}(z)>0$, (3.9) *has a unique solution $m(z)$ which has a positive imaginary part.*

The existence of a solution to equation (3.9) has already been proved in the proof of Theorem 3.4. To prove the uniqueness, rewrite equation (3.9) as

$$m=\int \frac{F^A(d\lambda)}{\lambda - z + xy}, \tag{3.12}$$

where

$$x = x(m) = \int \frac{\tau H(d\tau)}{1+\tau m}.$$

Suppose that the equation has two roots $m_1 \neq m_2$. Let $x_j = x(m_j)$, $j = 1, 2$. Then by (3.12), we have

$$\begin{aligned} m_1 - m_2 &= \int \frac{y(x_2-x_1)F^A(d\lambda)}{(\lambda - z + x_1 y)(\lambda - z + x_2 y)} \\ &= y(m_1-m_2)\int \frac{\tau^2 H(d\tau)}{(1+\tau m_1)(1+\tau m_2)} \int \frac{F^A(d\lambda)}{(\lambda - z + x_1 y)(\lambda - z + x_2 y)}. \end{aligned}$$

Finally, a contradiction can be derived by Hölder's inequality, as follows,

$$\begin{aligned} 1 &= y \int \frac{\tau^2 H(d\tau)}{(1+\tau m_1)(1+\tau m_2)} \int \frac{F^A(d\lambda)}{(\lambda - z + x_1 y)(\lambda - z + x_2 y)} \\ &\leq \Big(y \int \frac{\tau^2 H(d\tau)}{|1+\tau m_1|^2} \int \frac{F^A(d\lambda)}{|\lambda - z + x_1 y|^2}\Big)^{1/2} \Big(y \int \frac{\tau^2 H(d\tau)}{|1+\tau m_2|^2} \int \frac{F^A(d\lambda)}{|\lambda - z + x_2 y|^2}\Big)^{1/2} \\ &= \Big(y \int \frac{\tau^2 H(d\tau)}{|1+\tau m_1|^2} \frac{\mathrm{Im}(m_1)}{v + \mathrm{Im}(x_1) y}\Big)^{1/2} \Big(y \int \frac{\tau^2 H(d\tau)}{|1+\tau m_2|^2} \frac{\mathrm{Im}(m_2)}{v + \mathrm{Im}(x_2) y}\Big)^{1/2} < 1. \end{aligned}$$

Here, the last equality follows by comparing the imaginary part of equation (3.12) and the last inequality follows by observing that $\int \frac{F^A(d\lambda)}{|\lambda - z + x_j y|^2} = \frac{\mathrm{Im}(m_j)}{v - y\mathrm{Im}(x_j)}$ and $\mathrm{Im}(x_j) = \int \frac{\mathrm{Im}(m_j)\tau^2 H(d\tau)}{|1+\tau m_j|^2}$. The proof of the theorem is complete.

3.2. Convergence rates of spectral distributions

The problem of convergence rates of ESD's of LDRM had been open for decades since no suitable tools were found. As seen in Section 2, most important works were done by employing the MCT. Carleman's criterion guarantees convergence but does not give any rate. A breakthrough was made in recent work of Bai (1993a,b) in which Theorem 3.1 - Corollary 3.3 were proved and some convergence rates were established. Although these rates are still far from expected, some solid rates have been established and, more importantly, we have found a way to establish them. Bai, Miao and Tsay (1996a,b, 1997) further investigated the convergence rates of the ESD of large dimensional Wigner matrices.

3.2.1. Wigner matrix

In this section, we first introduce a result in Bai (1993a). Consider the model of Theorem 2.4 and assume that the entries of $\mathbf{W}$ above or on the diagonal are independent and satisfy

$$\begin{aligned} &\text{(i) } \mathrm{E}(w_{jk}) = 0, \text{ for all } 1 \le i \le j \le n;\\ &\text{(ii) } \mathrm{E}(|w_{jk}^2|) = 1, \text{ for all } 1 \le i < j \le n;\\ &\text{(iii) } \mathrm{E}(|w_{jj}^2|) = \sigma^2, \text{ for all } 1 \le j \le n\\ &\text{(iv) } \sup_n \max_{1\le i\le j\le n} \mathrm{E}(|w_{jk}^4|) \le M < \infty. \end{aligned} \tag{3.13}$$

Theorem 3.6. *Under the conditions in* (3.13)*, we have*

$$||\mathrm{E}F^{(1/\sqrt{n}\mathbf{W})} - F|| = O(n^{-1/4}), \tag{3.14}$$

where F is the semi-circular law with scale parameter 1.

Remark 3.3. The assertion (3.14) does not imply the complete convergence of $F^{(1/\sqrt{n}\mathbf{W})}$ to F. Here, we present a new result of Bai, Miao and Tsay (1997) in which a convergence rate in probability is established. Readers interested in the details of the proof of Theorem 3.6 are referred to Bai (1993a). Our purpose here is to illustrate how to use Theorem 3.1 - Corollary 3.3 to establish convergence rates of ESD's. Thus, we shall not pursue better rates through tedious arguments.

Theorem 3.7. *Under conditions* (i)-(iv) *in* (3.13)*, we have*

$$||F^{(1/\sqrt{n}\mathbf{W})} - F|| = O_p(n^{-1/4}). \tag{3.15}$$

Truncate the diagonal entries of $\mathbf{W}$ at $n^{1/8}$ and off-diagonal elements at $n^{1/3}$. Let $F_n^{(t)}$ denote the ESD of the truncated matrix. Then by Lemma 2.2 and condition (iv), we have

$$\begin{aligned} \mathrm{E}\|F_n - F_n^{(t)}\| &\le \frac{1}{n}\Big(\sum_{j\ne k} \mathrm{P}(|w_{jk}| \ge n^{1/3}) + \sum_k \mathrm{P}(|w_{kk}| \ge n^{1/8})\Big)\\ &\le \frac{Mn(n-1)n^{-4/3} + Mnn^{-1/2}}{n} \le 2Mn^{-1/3}. \end{aligned}$$

Centralize the off-diagonal elements of the truncated matrix, replace its diagonal elements by zero and denote the ESD of the resulting matrix by $F_n^{t,c}$. Then using Lemma 2.3, we obtain

$$\begin{aligned} L^3(F_n^{(t)} - F_n^{(t,c)}) &\le \frac{1}{n^2}\Big(\sum_{j\ne k} \mathrm{E}^2(|w_{jk}|I_{|w_{jk}|\ge n^{1/3}}) + \sum_k \mathrm{E}(|w_{kk}|^2 I_{|w_{kk}|\le n^{1/8}})\Big)\\ &\le \frac{M^2n(n-1)n^{-2} + nn^{1/4}}{n^2} \le 2n^{-3/4} \end{aligned}$$

for all large n. Therefore, to prove Theorem 3.6, we may make the additional assumptions that the diagonal elements of $\mathbf{W}$ are zero and the off-diagonal elements are bounded by $n^{-1/3}$. Then the conditions in Remark 2.8 are satisfied. Therefore, we have

$$\begin{aligned}
&\mathrm{E}\int_{|x|\geq 4}|F_n(x)-F(x)|dx\\
&\leq \int_4^{\infty}\Big(\mathrm{P}(\lambda_{\max}(n^{-1/2}\mathbf{W})\geq x)+\mathrm{P}(\lambda_{min}(n^{-1/2}\mathbf{W})\leq -x)\Big)dx=o(n^{-1}).
\end{aligned}$$

Recalling Theorem 3.2, we have for any $v>0$,

$$\begin{aligned}
&\mathrm{E}||F^{(1/\sqrt{n}\mathbf{W})}-F||\\
&\leq C\Big(\int_{|u|<16}\mathrm{E}(|m_n(z)-m(z)|)du+\mathrm{E}\int_{|x|\geq 4}|F_n(x)-F(x)|dx+v\Big)\\
&=C\int_{|u|<16}\mathrm{E}(|m_n(z)-m(z)|)du+O(n^{-1/4}),
\end{aligned}$$

if v is chosen to be $bn^{-1/4}$ for some $b>0$.

In Bai (1993a), it is proved that for the above chosen v,

$$\int|\mathrm{E}(m_n(z))-m(z)|du=O(v).$$

Thus, to prove (3.15), it is sufficient to prove

$$\int_{|u|<16}\mathrm{E}(|m_n(z)-\mathrm{E}(m_n(z))|)du=O(n^{-1/4}). \tag{3.16}$$

Define $\gamma_d=\mathrm{E}_d(m_n(z))-\mathrm{E}_{d-1}(m_n(z))$, $d=1,\ldots,n$, where E_d denotes the conditional expectation given the variables $\{w_{j,k},1\leq j\leq k\leq d\}$, with the convention that $\mathrm{E}_0=\mathrm{E}$. Note that $\{\gamma_1,\ldots,\gamma_n\}$ forms a martingale difference sequence and

$$m_n(z)-\mathrm{E}(m_n(z))=\frac{1}{n}\sum_{d=1}^{n}\gamma_d.$$

By noticing $|\gamma_k|\leq 2/v$ and the orthogonality of martingale differences, we get

$$\begin{aligned}
\mathrm{E}|m_n(z)-\mathrm{E}(m_n(z))|&\leq \mathrm{E}^{1/2}|m_n(z)-\mathrm{E}(m_n(z))|^2\\
&=\frac{1}{n}\left(\sum_{d=1}^{n}\mathrm{E}(|\gamma_d|^2)\right)^{1/2}\\
&\leq 2v^{-1}n^{-1/2}=O(n^{-1/4}).
\end{aligned}$$

The proof of the theorem is complete.

In Bai, Miao and Tsay (1996a,b), the convergence rate of Wigner matrices is investigated further. The following results are established in the first of these works.

Theorem 3.8. *Suppose that the diagonal entries of* $\mathbf{W}$ *are i.i.d. with mean zero and finite sixth moment and that the elements above the diagonal are i.i.d. with mean zero, variance* 1 *and finite eighth moment. Then the following results are true:*

$$\|\mathrm{E}F_n - F\| = O(n^{-1/2})$$

and

$$\|F_n - F\| = O_p(n^{-2/5}).$$

If we further assume that the entries of $\mathbf{W}$ *have finite moments of all orders, then for any* $\varepsilon > 0$,

$$\|F_n - F\| = o_{a.s.}(n^{-2/5+\varepsilon}).$$

In Bai, Miao and Tsay (1996b), the convergence rate of the expected ESD of $\mathbf{W}$ is improved to $O(n^{-1/3})$ under the conditions of Theorem 3.6.

3.2.2. Sample covariance matrix

Assume the following conditions are true.

(i) $\mathrm{E}(x_{jk}) = 0$, $\mathrm{E}(|x_{jk}^2|) = 1$, for all j, k, n,

(ii) $\sup_n \sup_{j,k} \mathrm{E}|x_{jk}^4|I_{(|x_{jk}|\geq M)} \to 0$, as $M \to \infty$. (3.17)

In Bai (1993b), the following theorems are proved.

Theorem 3.9. *Under the assumptions in* (3.17)*, for* $0 < \theta < \Theta < 1$ *or* $1 < \theta < \Theta < \infty$,

$$\sup_{y_p \in (\theta,\Theta)} \|\mathrm{E}F^{\mathbf{S}} - F_{y_p}\| = O(n^{-1/4}), \tag{3.18}$$

where $y_p = p/n$ *and* F_{y_p} *is defined in Theorem* 2.19.

Theorem 3.10. *Under the assumptions in* (3.17)*, for any* $0 < \varepsilon < 1$,

$$\sup_{y_p \in (1-\varepsilon,1+\varepsilon)} ||\mathrm{E}F^{\mathbf{S}} - F_{y_p}|| = O(n^{-5/48}). \tag{3.19}$$

By the same approach as in the proof of Theorem 3.8, Bai, Miao and Tsay (1996a) also generalized the results of Theorems 3.9 and 3.10 to the following theorem.

Theorem 3.11. *Under the assumptions in* (3.17)*, the conclusions in Theorems* 3.9 *and* 3.10 *can be improved to*

$$\sup_{y_p \in (\theta,\Theta)} ||F^{\mathbf{S}} - F_{y_p}|| = O_p(n^{-1/4})$$

and

$$\sup_{y_p \in (1-\varepsilon, 1+\varepsilon)} ||F^{\mathbf{S}} - F_{y_p}|| = O_p(n^{-5/48}).$$

4. Circular Law - Non-Hermitian Matrices

In this section, we consider a kind of non-Hermitian matrix. Let $\mathbf{Q} = n^{-1/2}(x_{jk})$ be an $n \times n$ complex matrix with i.i.d. entries x_{jk} of mean zero and variance 1. The eigenvalues of $\mathbf{Q}$ are complex and thus the ESD of $\mathbf{Q}$, denoted by $F_n(x, y)$, is defined in the complex plane. Since the early 1950's, it has been conjectured that $F_n(x, y)$ tends to the uniform distribution over the unit disc in the complex plane, called the circular law. The major difficulty is that the major tools introduced in the previous two sections do not apply to non-Hermitian matrices.

Ginibre (1965) found the density of the eigenvalues of a matrix of i.i.d. complex $N(0, 1)$ entries to be

$$c \prod_{j \neq k} |\lambda_j - \lambda_k|^2 \exp\{-\frac{1}{2}\sum_{k=1}^{n} |\lambda_k|^2\}.$$

Based on this result, Mehta (1991) proved the circular law when the entries are i.i.d. complex normally distributed. Hwang (1986) reported that this result was also proved in an unpublished paper of Silverstein by the same approach.

Girko (1984a,b) presented a proof of the circular law under the condition that the entries have bounded densities on the complex plane and finite $(4 + \varepsilon)$th moments. Since they were published, many have tried to understand his mathematical arguments without success. The problem was considered open until Bai (1997) proved the following.

Theorem 4.1. *Suppose that the entries have finite* $(4 + \varepsilon)$*th moments, and that the joint distribution of the real and imaginary parts of the entries, or the conditional distribution of the real part given the imaginary part, has a uniformly bounded density. Then the circular law holds.*

Remark 4.1. The second part of Theorem 4.1 covers real random matrices. In this case, the joint distribution of the real and imaginary parts of the entries does not have a density in the complex plane. However, when the entries are real and have a bounded density, the real and imaginary parts are independent and hence the condition in the second part of Theorem 4.1 is satisfied. By considering the matrix $e^{i\theta}\mathbf{X}$, we can extend the density condition in the second part of Theorem 4.1 to: *the conditional density of* $\mathrm{Re}(x_{jk})\cos(\theta) - \mathrm{Im}(x_{jk})\sin(\theta)$ *given* $\mathrm{Re}(x_{jk})\sin(\theta) + \mathrm{Im}(x_{jk})\cos(\theta)$ *is bounded.*

Although Girko's arguments are hard to understand, or even deficient, he provided the following idea. Let $F_n(x, y)$ denote the ESD of $n^{-1/2}\mathbf{X}$, and $\nu_n(x, z)$

denote the ESD of the Hermitian matrix $\mathbf{H} = \mathbf{H}_n(z) = (n^{-1/2}\mathbf{X} - z\mathbf{I})(n^{-1/2}\mathbf{X} - z\mathbf{I})^*$ for given $z = s + it$.

Lemma 4.2. (Girko) *For* $uv \neq 0$,

$$\int\int e^{iux+iyv} F_n(dx, dy) = \frac{u^2+v^2}{4\pi iu} \int\int e^{isu+ivt} g_n(s,t) dt ds, \tag{4.1}$$

and

$$\int\int e^{iux+iyv} F_{\text{cir}}(dx, dy) = \frac{u^2+v^2}{4\pi iu} \int\int e^{isu+ivt} g(s,t) dt ds, \tag{4.2}$$

where

$$g_n(s,t) = \frac{\partial}{\partial s} \int_0^\infty \log x \nu_n(dx, z),$$

F_{cir} *is the uniform distribution over the unit disc in the complex plane, and* $g(s,t) = 2s$ *or* $2s/|z|^2$ *in accordance with* $|z| < 1$ *or not.*

Making use of the formula that for all $uv \neq 0$,

$$\frac{u^2+v^2}{2iu\pi} \int \left[\int \frac{s}{s^2+t^2} e^{ius+ivt} dt \right] ds = 1,$$

we obtain

$$\begin{aligned}
&\int\int e^{iux+ivy} F_n(dx, dy) \\
&= \frac{u^2+v^2}{2iu\pi} \int\int \frac{1}{n} \sum_{k=1}^n \frac{s}{s^2+t^2} e^{ius+ivt+iu\text{Re}(\lambda_k)+iv\text{Im}(\lambda_k)} dt ds \\
&= \frac{u^2+v^2}{4iu\pi} \int\int \frac{1}{n} \sum_{k=1}^n \frac{2(s-\text{Re}(\lambda_k))}{(s-\text{Re}(\lambda_k))^2 + (t-\text{Im}(\lambda_k))^2} e^{ius+ivt} dt ds \\
&= \frac{u^2+v^2}{4iu\pi} \int\int \frac{\partial}{\partial s} \frac{1}{n} \sum_{k=1}^n \log(|z-\lambda_k|^2) e^{ius+ivt} dt ds \\
&= \frac{u^2+v^2}{4iu\pi} \int\int \left[\frac{\partial}{\partial s} \int_0^\infty \log x \nu_n(dx, z) \right] e^{ius+ivt} dt ds.
\end{aligned} \tag{4.3}$$

Here, we have used the fact that $\prod_{k=1}^n |z - \lambda_k|^2 = \det(\mathbf{H})$. The proof of the first assertion of Lemma 4.2 is complete. The second assertion follows from the Green Formula.

Under the condition that the entries have finite $(4+\varepsilon)$th moments, it can be shown that, as mentioned in Subsection 2.2.2, the upper limit of the maximum absolute value of the eigenvalues of $n^{-1/2}\mathbf{X}$ is less than the maximum singular value, which tends to 2. Thus the distribution family $\{F_n(x,y)\}$ is tight. Hence going along some subsequence of integers, F_n and $\nu_n(x,z)$ tend to limits μ and ν respectively. It seems the circular law follows by making limit in (4.1) and getting (4.2) with μ and ν substituting F_{cir} and the ν defined by the circular law. However, there is no justifcation for passing the limit procedure $\nu_n \to \nu$ through the 3-fold integration since the outside integral range in (4.3) is the whole plane and the integrand of the inner integral is unbounded. To overcome the first difficulty, we need to reduce the integral range.

Let $T = \{z;\ |s| < A,\ |t| < A^2, |1 - |z|| > \varepsilon\}$.

Lemma 4.3. *For any $A > 0$ and $\varepsilon > 0$, with probability* 1,

$$\left|\int\int_{T^c} e^{isu+ivt} g_n(s,t)dtds\right| = O(A^{-1} + \varepsilon) \quad \textit{uniformly in } n.$$

The same is true if g_n is replaced by g, where g is defined in Lemma 4.2.

By the lemma and integration by parts, the problem is reduced to showing that

$$\int\int_T \left|\int_0^\infty \log x(\nu_n(dx,z) - \nu(dx,z))\right| dtds \to 0 \quad a.s. \tag{4.4}$$

Since $z \in T$ and the norm of $\mathbf{Q}$ is bounded with probability 1, the support of $\nu_n(x,z)$ is bounded above by, say, M. Therefore, it is not a problem when dealing with the upper limit of the inner integral. However, since $\log x$ is not bounded at zero, (4.4) could not follow from $\nu_n \to \nu$. To overcome this difficulty, we estimate the convergence rate of $\nu_n - \nu$ and prove the following lemma.

Lemma 4.4. *Under the conditions of Theorem* 4.1*, we have*

$$\sup_{z\in T} ||\nu_n(\cdot,z) - \nu(\cdot,z)|| = o(n^{-\beta}),\ \ a.s.,$$

where $\beta > 0$ depends on ε (in the moment condition) only.

Let $\varepsilon_n = e^{-n^\beta}$. Then by Lemma 4.4,

$$\sup_{z\in T}\left|\int_{\varepsilon_n}^\infty \log x(\nu_n(dx,z) - \nu(dx,z))\right| \le n^\beta M \sup_{z\in T} ||\nu_n(\cdot,z) - \nu(\cdot,z)|| = o(1), \text{a.s.}$$

It remains to show that

$$\int\int_T \int_0^{\varepsilon_n} \log x\nu_n(dx,z)dtds \to 0 \ \ \text{a.s.} \tag{4.5}$$

$$\int\int_T \int_0^{\varepsilon_n} \log x\nu(dx,z)dtds \to 0 \ \ \text{a.s.}$$

The most difficult part is the proof of (4.5). For details, see Bai (1997).

5. Applications

In this section, we introduce some recent applications in multivariate statistical inference and signal processing. The examples discussed reveal that when the dimension of the data or parameters to be estimated is "very high", it causes non-negligible errors in many traditional multivariate statistical methods. Here, "very high" does not mean "incredibly" high, but "fairly" high. As simulation results for problems in the following sub-sections show (see cited papers), when the ratio of the degrees of freedom to dimension is less than 5, the non-exact test significantly beats the traditional T^2 in a two-sample problem (see Bai and Saranadasa (1996) for details); in the detection of the number of signals in a multivariate signal processing problem, when the number of sensors is greater than 10, the traditional MUSIC (MUltivariate SIgnal Classification) approach performs poorly, even when the sample size is as large as 1000. Such a phenomenon has been found in many different areas. In a normality test, say, the simplified W'-test beats Shapiro's W-test for most popular alternatives, although the latter is constructed by the Markov-Gaussian method, seemingly more reasonable than the usual least squares method. I was also told that when the number of regression coefficients in a multivariate regression problem is more than 6, the estimation becomes worse, and that when the number of parameters in a structured covariance matrix is more than 4, the estimates have serious errors. In applied time series analysis, models with orders greater than 6 (p in AR model, q in MA and $p+q$ in ARMA) are seldom considered. All these tell us that one has to be careful when dealing with high-dimensional data or a large number of parameters.

5.1. Non-exact test for the two-sample problem

Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ are random samples from two populations with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and a common covariance matrix Σ. Our problem is to test the hypothesis $H: \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against $K: \ \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. The classical approach uses the Hotelling test (or T^2-test), with

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{\mathbf{x}} - \overline{\mathbf{y}})'\mathbf{A}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}),$$

$\overline{\mathbf{x}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_i$, $\overline{\mathbf{y}} = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{y}_i$ and

$$\mathbf{A} = \frac{1}{n_1 + n_2 - 2}\Big(\sum_{i=1}^{n_1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' + \sum_{i=1}^{n_2}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'\Big).$$

The T^2 test has lots of good properties, but it is not well defined when the degrees of freedom $(n_1 + n_2 - 2)$ is less than the dimension (p) of the data.

As a remedy, Dempster (1959) proposed the so-called *non-exact test* (NET) by using the chi-square approximation technique. In recent research of Bai and Saranadasa (1996), it is found that Dempster's NET is also much more powerful than the T^2 test in many general situations when T^2 is well defined. One difficulty in computing Dempster's test statistic is the construction of a high dimensional orthogonal matrix and the other is the estimation of the degrees of freedom of the chi-square approximation. Bai and Saranadasa (1996) proposed a new test, the asymptotic normal test (ANT), in which the test statistic is based on $\|\overline{\mathbf{x}} - \overline{\mathbf{y}}\|^2$, normalized by consistent estimators of its mean and variance. It is known that ANT is asymptotically equivalent to NET, and simulations show that ANT is slightly more powerful than NET. It is easy to show that the type I errors for both NET and ANT tend to the prechosen level of the test. Simulation results show that NET and ANT gain a great amount of power with a slight loss of the exactness of the type I error. Note that *non-exact* does not mean that the error is larger.

Now, let us analyze why this happens. Under the normality assumption, if Σ were known, then the "most powerful test statistic" should be $(\overline{\mathbf{x}}-\overline{\mathbf{y}})'\Sigma^{-1}(\overline{\mathbf{x}}-\overline{\mathbf{y}})$. Since Σ is actually unknown, the matrix $\mathbf{A}$ plays the role of an estimator of Σ. Then there is the problem of how close $\mathbf{A}^{-1}$ is to Σ^{-1}. The matrix $\mathbf{A}^{-1}$ can be rewritten in the form $\Sigma^{-1/2}\mathbf{S}^{-1}\Sigma^{-1/2}$, where $\mathbf{S}$ is defined in Subsection 2.1.2, with $n = n_1 + n_2 - 2$. The approximation is good if $\mathbf{S}^{-1}$ is close to $\mathbf{I}$. Unfortunately, this is not really the case. For example, when $p/n = 0.25$, the ratio of the largest eigenvalue of $\mathbf{S}^{-1}$ to the smallest can be as large as 9. Even when p/n is as small as 0.01, the ratio can be as large as 1.493. This shows that it is practically impossible to get a "good" estimate of the inverse covariance matrix. In other words, if the ratio of the largest to the smallest eigenvalues of the population covariance matrix is not larger than $(\sqrt{n}+\sqrt{p})^2/(\sqrt{n}-\sqrt{p})^2$ (e.g. 9 for $p/n = 0.25$ and 1.493 for $p/n = 0.01$), NET or ANT give a better test than T^2.

A similar but simpler case is the one-sample problem. As in Bai and Saranadasa (1996), it can be shown that NET and ANT are better than the T^2 test. This phenomenon happens in many statistical inference problems, such as large contingency tables, MANOVA, discretized density estimation, linear models with large number of parameters and the Error in Variable Models. Once the dimension of the parameter is large, the performance of the classical estimators become poor and corrections may be needed.

5.2. Multivariate discrimination analysis

Suppose that $\mathbf{x}$ is a sample drawn from one of two populations with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and a common covariance matrix Σ. Our problem is to classify the present sample $\mathbf{x}$ into one of the two populations. If $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and Σ are known, then the best discriminant function is $d = (\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2))'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$, i.e., assign $\mathbf{x}$ to Population 1 if $d > 0$.

When both the mean vectors and the covariance matrix are unknown, assume training samples $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ and $\mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$ from the two populations are available. Then we can substitute the MLE $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ and $\mathbf{A}$ of the mean vectors and covariance matrix into the discriminant function. Obviously, this is impossible if $n = n_1 + n_2 - 2 < p$. The problem is again whether this criterion has the smallest misclassification probability when p is large. If not, what discrimination criterion is better. Based on the same discussion in the last subsection, one may guess that the criterion $d = (\mathbf{x} - \frac{1}{2}(\overline{\mathbf{x}} + \overline{\mathbf{y}}))'(\overline{\mathbf{x}} - \overline{\mathbf{y}})$ should be better. Using the LSD of a large sample covariance matrix, this was theoretically proved in Saranadasa (1993). Simulation results presented in his paper strongly support the theoretical results, even for moderate n and p.

5.3. Detection of the number of signals

Consider the model

$$\mathbf{y}_j = \mathbf{A}\mathbf{s}_j + \mathbf{n}_j, \quad j = 1, \ldots, N,$$

where $\mathbf{y}_j$ is a $p \times 1$ complex vector of observations collected from p sensors, $\mathbf{s}_j$ a $q \times 1$ complex vector of unobservable signals emitted from q targets, $\mathbf{A}$ is an unknown $p \times q$ matrix whose columns are called the distance-direction vectors and $\mathbf{n}_j$ represents the noise generated by the sensors, usually assumed to be white. Usually, in detecting the number of signals (for non-coherent models), $\mathbf{A}$ is assumed to be of full rank and the number q is assumed to be less than p. In the estimation of DOA (Direction Of Arrivals), the (j, k)th element of $\mathbf{A}$ is assumed to be $r_k \exp(-2\pi d(j-1)\omega_0 \sin(\theta_k)/c)$, where r_k is the complex amplitude determined by the distance from the kth target to the jth sensor, d the spatial distance between adjacent sensors, ω_0 the central frequency, c the speed of light and θ_k the angle between the line of the sensors and the line from the jth sensor to the kth target, called the DOA. The most important problems are the detection of the number q of signals and the estimation of the DOA. In this section, we only consider the detection of the number of signals.

All techniques for solving the problem are based on the following:

$$\Sigma_{\mathbf{y}} = \mathbf{A}\boldsymbol{\Psi}\mathbf{A}^* + \sigma^2\mathbf{I},$$

where $\boldsymbol{\Psi}$ $(q \times q)$ is the covariance matrix of the signals. Denote the eigenvalues of $\Sigma_{\mathbf{y}}$ by $\lambda_1 \geq \cdots \geq \lambda_q > \lambda_{q+1} = \cdots = \lambda_p = \sigma^2$. This means that the multiplicity of the smallest eigenvalues σ^2 is $p - q$ and there is a gap between λ_q and λ_{q+1}.

Since the signals and noise have zero means, one can use $\widehat{\Sigma}_N = \frac{1}{N}\sum_{j=1}^{N} \mathbf{y}_j\mathbf{y}_j^*$ as an estimator of $\Sigma_{\mathbf{y}}$, and then compare a few of the smallest eigenvalues of $\widehat{\Sigma}_N$ to estimate the number of signals q. In the literature, AIC, BIC and GIC criteria are used to detect the number of signals. However, when p is large, the problem is then how big the gap between the qth and $(q+1)$-st largest eigenvalues of $\widehat{\Sigma}_N$ should be, so that q can be correctly detected by these criteria. Simulations in the literature usually take q to be 2 or 3 and p to be 4 or 5. Once $p = 10$ and $SNR = 0db$ (SNR (Signal to noise ratio) is defined as ten times the logarithm of the ratio of the variance of the signal (the kth component of $\mathbf{As}_1$) to the variance of the noise (the kth component of $\mathbf{n}_1$)), no criterion works well unless N is larger than 1000 (i.e. $y \sim 0.01$). Unreasonably, if we drop half of the data (i.e., reduce p to 5), the simulation results become good even for $n = 300$ or 400.

From the theory of LDRM, in the support of the LSD of $\widehat{\Sigma}_N$, there may be a gap at the $(1-q/p)$th quantile or the gap may disappear, which depends on the original gap and the ratio $y = \lim p/N$ in a complicated manner. Some work was done in Silverstein and Combettes (1992). Their simulation results show that when the gap exists in the support of the LSD, the exact number q (not only the ratio q/p) can be exactly estimated for all large N. More precisely, suppose $p = p_N$ and $q = q_N$ tend to ∞ proportionally to N, then $\mathrm{P}(\hat{q}_N \neq q, i.o.) = 0$. Work on this is being done by Silverstein and myself (see Bai and Silverstein (1998)).

6. Unsolved Problems

6.1. Limiting spectral distributions

6.1.1. Existence of the LSD

Nothing is known about the existence of the LSD's of the following three matrices

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ x_2 & x_1 & \ddots & x_{n-1} \\ \vdots & \ddots & \ddots & \vdots \\ x_n & x_{n-1} & \cdots & x_1 \end{pmatrix}, \qquad \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ x_2 & x_3 & \cdots & x_{n+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_{n+1} & \cdots & x_{2n-1} \end{pmatrix}$$

and

$$\begin{pmatrix} -\sum_{i=2}^{n} x_{1i} & x_{12} & \cdots & x_{1n} \\ x_{21} & -\sum_{i=1,\neq 2}^{n} x_{2i} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & -\sum_{i=1}^{n-1} x_{ni} \end{pmatrix},$$

where, in the first two matrices, x_j's are i.i.d. real random variables and in the third matrix, $x_{jk} = x_{kj}$, $j < k$, are i.i.d. real random variables. Consider the three matrices as limiting distributions of the form $\sqrt{n}(\mathbf{A}_n - \mathbf{A})$: the first is for the autocovariance matrix in time series analysis, the second is for the information matrix in a polynomial regression model and the third is for the derivative of a transition matrix in a Markov process.

6.1.2. Explicit forms of LSD

The only known explicit forms of densities of LSD's of LDRM are those of the semi-circular law, the circular law, the Marčenko-Pastur law and the Multivariate F-matrix. As shown in Theorem 3.4, there are a large class of random matrices whose LSD's exist but for which no explicit forms are known. It is of interest to find more explicit forms of the densities of the LSD's.

6.2. Limits of extreme eigenvalues

These are known for Wigner and sample covariance matrices. Nothing is known for multivariate F matrices.

As mentioned in Section 5.3, it is very interesting that there are no eigenvalues at all in the gap of the support (this is called the separation problem) of the LSD. More precisely, suppose that $q_N/N \to z$ and $p_N/N \to y$ with $0 < z < y < 1$, and $\lambda_{q_N} \to c$, $\lambda_{q_N+1} \to d$ with $d < c$. Under certain conditions, we conjecture that $\ell_{q_N} \to g_{z,y}(c)$ and $\ell_{q_N+1} \to g_{z,y}(d)$, where λ_{q_N}, λ_{q_N+1}, ℓ_{q_N} and ℓ_{q_N+1} are the q_Nth and (q_N+1)-st largest eigenvalues of Σ and $\mathbf{S}\Sigma$, respectively, and $g_{z,y}(c) > g_{z,y}(d)$ are the upper and lower bounds of the $(1-z)$-quantile of the LSD of $\mathbf{S}\Sigma$.

Remark 6.1. After this paper was written, the above mentioned problem has been partially solved in Bai and Silverstein (1998). For details, see Silverstein's discussion following this paper.

6.3. Convergence rates of spectral distributions

The only known results are introduced in Subsection 3.2. For Wigner and sample covariance matrices, some convergence rates of ESD's are given in Bai (1993a,b), Bai, Miao and Tsay (1996a,b, 1997) and the present paper. Of more interest is the rates of a.s. or in probability convergence. It is also of interest is to find the ideal convergence rates (the conjectured rates are of the order $O(1/n)$ or at least $O(1/\sqrt{n})$). Furthermore, nothing is known about other matrices.

6.4. Second order convergence

6.4.1. Second order convergence of spectral distributions

Of course, the convergence rates should be determined first. Suppose that the exact rate is found to be α_n. It is reasonable to conjecture that $\alpha_n^{-1}(F_n(x)-F(x))$ should tend to a limiting stochastic process. Based on this, it may be possible to find limiting distributions of statistics which are functionals of the ESD. Then statistical inference, such as testing of hypothesis and confidence intervals, can be performed.

6.4.2. Second order convergence of extreme eigenvalues

In Subsection 2.2, limits of extreme eigenvalues of some random matrices are presented. As mentioned in the last subsubsection, it is important to find the limiting distribution of $\alpha_n^{-1}(\ell_{\rm extr} - \lambda_{\rm lim})$, where $\ell_{\rm extr}$ is the extreme eigenvalue and $\lambda_{\rm lim}$ is the limit of $\ell_{\rm extr}$. The normalizing constant α_n may be the same as, or different from, that for the corresponding ESD's. For example, for Wigner and sample covariance matrices with $y \neq 1$, the conjectured α_n is $\frac{1}{n}$, but for sample covariance matrices with $p = n$, the conjectured normalizing constant for the smallest eigenvalue of $\mathbf{S}$ is $1/n^2$. The smallest eigenvalue when $p = n$ is related to the condition number (the square-root of the ratio of the largest to the smallest eigenvalues of $\mathbf{S}$), important in numerical computation of linear equations. Reference is made to Edelman (1992).

6.4.3. Second order convergence of eigenvectors

Some results on the eigenvectors of large-dimensional sample covariance matrices were established in the literature and introduced in Subsection 2.3. A straightforward problem is to extend these results to other kinds of random matrices. Another problem is whether there are other ways to describe the similarity between the eigenmatrix and Haar measure.

6.5. Circular law

The conjectured condition for guaranteeing the circular law is finite second moment only, at least for the i.i.d. case. In addition to the difficulty of estimating (4.5), there are no similar results to Lemmas 2.2, 2.3, 2.6 and 2.7, so we cannot truncate the variable at $\sqrt{n}$ under the existence of the second moment of the underlying distributions.

Acknowledgement

The author would like to thank Professor J. W. Silverstein for indicating that the eigenvalues of the matrix with elements $i, -i$ and 0 above, below and on the diagonal are given by $\cot(\pi(2k-1)/2n)$, $k = 1, \ldots, n$. This knowledge plays a key role in dealing with the expected imaginary parts of the entries of a Wigner matrix in Theorems 2.1 and 2.12.

References

Arnold, L. (1967). On the asymptotic distribution of the eigenvalues of random matrices. *J. Math. Anal. Appl.* **20**, 262-268.

Arnold, L. (1971). On Wigner's semicircle law for the eigenvalues of random matrices. *Z. Wahrsch. Verw. Gebiete* **19**, 191-198.

Bai, Z. D. (1993a). Convergence rate of expected spectral distributions of large random matrices. Part I. Wigner Matrices. *Ann. Probab.* **21**, 625-648.

Bai, Z. D. (1993b). Convergence rate of expected spectral distributions of large random matrices. Part II. Sample Covariance Matrices. *Ann. Probab.* **21**, 649-672.

Bai, Z. D. (1997). Circular law. *Ann. Probab.* **25**, 494-529.

Bai, Z. D., Miao, B. Q. and Tsay, J. (1996a). Convergence rates of the spectral distributions of large Wigner matrices. Submitted.

Bai, Z. D., Miao, B. Q. and Tsay, J. (1996b). Remarks on the convergence rate of the spectral distributions of Wigner matrices. Submitted.

Bai, Z. D., Miao, B. Q. and Tsay, J. (1997). A note on the convergence rate of the spectral distributions of large random matrices. *Statist. Probab. Lett.* **34**, 95-102.

Bai, Z. D. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statist. Sinica* **6**, 311-329.

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

Bai, Z. D., Silverstein, Jack W. and Yin, Y. Q. (1988). A note on the largest eigenvalue of a large dimensional sample covariance matrix. *J. Multivariate Anal.* **26**, 166-168.

Bai, Z. D. and Yin, Y. Q. (1986). Limiting behavior of the norm of products of random matrices and two problems of Geman-Hwang. *Probab. Theory Related Fields* **73**, 555-569.

Bai, Z. D. and Yin, Y. Q. (1988a). Convergence to the semicircle law. *Ann. Probab.* **16**, 863-875.

Bai, Z. D. and Yin, Y. Q. (1988b). Necessary and sufficient conditions for the almost sure convergence of the largest eigenvalue of a Wigner matrix. *Ann. Probab.* **16**, 1729-1741.

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

Bai, Z. D., Yin, Y. Q. and Krishnaiah, P. R. (1986). On LSD' of product of two random matrices when the underlying distribution is isotropic. *J. Multivariate Anal.* **19**, 189-200.

Bai, Z. D., Yin, Y. Q. and Krishnaiah, P. R. (1987). On the limiting empirical distribution function of the eigenvalues of a multivariate F matrix. *Theory Probab. Appl.* **32**, 490-500.

Edelman, A. (1992). On the distribution of a scaled condition number. *Math. Comp.* **58**, 185-190.

Edelman, A. (1997). The circular law and the probability that a random matrix has k real eigenvalues. *J. Multivariate Anal.* **60**, 188-202.

Geman, S. (1980). A limit theorem for the norm of random matrices. *Ann. Probab.* **8**, 252-261.

Geman, S. (1986). The spectral radius of large random matrices. *Ann. Probab.* **14**, 1318-1328.

Ginibre, J (1965). Statistical ensembles of complex, quaterion and real matrices. *J. Math. Phys.* **6**, 440-449.

Girko, V. L. (1984a). Circle law. *Theory Probab. Appl.* **4**, 694-706.

Girko, V. L. (1984b). On the circle law. *Theory Probab. Math. Statist.* **28**, 15-23.

Girko, V. L. (1990). *Theory of Random Determinants.* Kluwer Academic Publishers, Dordrecht-Boston-London.

Girko, V., Kirsch, W. and Kutzelnigg, A. (1994). A necessary and sufficient condition for the semicircular law. *Random Oper. Stoch. Equ.* **2**, 195-202.

Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables.* Addison-Wesley, Reading.

Grenander, Ulf (1963). *Probabilities on Algebraic Structures.* John Wiley, New York-London.

Grenander, Ulf and Silverstein, Jack W. (1977). Spectral analysis of networks with random topologies. *SIAM J. Appl. Math.* **32**, 499-519.

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

Hwang, C. R. (1986). A brief survey on the spectral radius and the spectral distribution of large dimensional random matrices with i.i.d. entries. *Random Matrices and Their Applications*, Contemporary Mathematics **50**, 145-152, AMS, Providence.

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

Loève, M. (1977). *Probability Theory.* 4th edition. Springer-Verlag, New York.

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

Mehta, M. L. (1991). *Random Matrices.* Academic Press, New York.

Pastur, L. A. (1972). On the spectrum of random matrices. *Teoret. Mat. Fiz.* **10**, 102-112, (Teoret. Mat. Phys. **10**, 67-74).

Pastur, L. A. (1973). Spectra of random self-adjoint operators. *Uspekhi Mat. Nauk.* **28**, 4-63, (Russian Math. Surveys **28**, 1-67).

Prohorov, Ju. V. (1968). The extension of S. N. Bernstein's inequalities to a multi-dimensional case. (Russian) *Teor. Verojatnost. i Primenen.* **13**, 266-274.

Rao, C. R. (1976). *Linear Statistical Inference and Its Applications.* 2nd edition. John Wiley, New York.

Saranadasa, H. (1993). Asymptotic expansion of the misclassification probabilities of D- and A-criteria for discrimination from two high dimensional populations using the theory of large dimensional random matrices *J. Multivariate Anal.* **46**, 154-174.

Silverstein, J. W. (1979). On the randomness of eigenvectors generated from networks with random topologies. *SIAM J. Appl. Math.* **37**, 235-245.

Silverstein, J. W. (1981). Describing the behavior of eigenvectors of random matrices using sequences of measures on orthogonal groups. *SIAM J. Math. Anal.* **12**, 174-281.

Silverstein, J. W. (1984a). Comments on a result of Yin, Bai and Krishnaiah for large dimensional multivariate F matrices. *J. Multivariate Anal.* **15**, 408-409.

Silverstein, J. W. (1984b). Some limit theorems on the eigenvectors of large dimensional sample covariance matrices. *J. Multivariate Anal.* **15**, 295-324.

Silverstein, J. W. (1985a). The limiting eigenvalue distribution of a multivariate F matrix. *SIAM J. Appl. Math.* **16**, 641-646.

Silverstein, J. W. (1985b). The smallest eigenvalue of a large dimensional Wishart Matrix. *Ann. Probab.* **13**, 1364-1368.

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

Silverstein, J. W.(1989b). On the weak limit of the largest eigenvalue of a large dimensional sample covariance matrix *J. Multivariate Anal.* **30**, 307-311.

Silverstein, J. W. (1990). Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *Ann. Probab.* **18**, 1174-1194.

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

Silverstein J. W. and Bai, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *J. Multivariate Anal.* **54**, 175-192.

Silverstein, W. J. and Choi, S. I. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.* **54**, 295-309.

Silverstein, J. W. and Combettes, P. L. (1992). Signal detection via spectral theory of large dimensional random matrices. *IEEE ASSP* **40**, 2100-2104.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probab.* **6**, 1-18.
Wachter, K. W. (1980). The limiting empirical measure of multiple discriminant ratios. *Ann. Statist.* **8**, 937-957.
Wigner, E. P. (1955). Characteristic vectors bordered matrices with infinite dimensions. *Ann. Math.* **62**, 548-564.
Wigner, E. P. (1958). On the distributions of the roots of certain symmetric matrices. *Ann. Math.* **67**, 325-327.
Yin, Y. Q. (1986). LSD' for a class of random matrices. *J. Multivariate Anal.* **20**, 50-68.
Yin, Y. Q., Bai, Z. D. and Krishnaiah, P. R. (1983). Limiting behavior of the eigenvalues of a multivariate F matrix *J. Multivariate Anal.* **13**, 508-516.
Yin, Y. Q., Bai, Z. D. and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probab. Theory Related Fields* **78**, 509-521.
Yin, Y. Q. and Krishnaiah, P. R. (1983). A limit theorem for the eigenvalues of product of two random matrices. *J. Multivariate Anal.* **13**, 489-507.
Yin, Y. Q. and Krishnaiah, P. R. (1985). Limit theorem for the eigenvalues of the sample covariance matrix when the underlying distribution is isotropic. *Theory Probab. Appl.* **30**, 861-867.

Department of Statistics and Applied Probability, National University of Singapore, Singapore 119260.
E-mail: matbaizd@leonis.nus.edu.sg

(Received January 1996; accepted March 1999)

COMMENT: SPECTRAL ANALYSIS OF RANDOM MATRICES USING THE REPLICA METHOD

G. J. Rodgers

Brunel University

Abstract: In this discussion paper, we give a brief review of the replica method applied to random matrices, and in particular to their spectral analysis. We illustrate the method by calculating the eigenvalue spectrum of the real random matrix ensemble describing the Hopfield model of autoassociative memory.

Key words and phrases: Random matrices, replica method, spectral analysis.

1. Introduction

In Bai (1999), the author reviews the theory of random matrices from the mathematical physics literature. In contrast to this rigorous analysis of spectral theory, there have been parallel, non-rigorous, developments in the theoretical physics literature. Here the replica method, and to a lesser extent super-symmetric methods, have been used to analyse the spectral properties of a variety of random matrices of interest to theoretical physicists. These matrices have applications in, for instance, random magnet theory, neural network theory and the conductor/insulator transition. In the present discussion we briefly review the work using the replica method. We then illustrate the use of this method by using it, for the first time, to obtain the spectral distribution of the sample covariance matrix. This problem is considered in Section 2.1.2 of Bai (1999) using a completely different approach.

The replica method was introduced by Edwards (1970) to study a polymer physics problem. It was first applied to a matrix model by Edwards and Jones (1976) who used it to obtain the Wigner semi-circular distribution for the spectrum of a random matrix with Gaussian distributed entries. Since then it was applied by Rodgers and Bray (1988) and Bray and Rodgers (1988) to obtain the spectral distribution of two different classes of sparse random matrices. Later, Sommers, Crisanti, Sompolinsky and Stein (1988) used an electrostatic method, which nevertheless relied on the replica method to demonstrate an assumption, to obtain the average eigenvalue distribution of random asymmetric matrices. Some of these approaches are analogous to the super-symmetric technique used on sparse random matrices by Rodgers and DeDominicis (1990) and Mirlin and Fyodorov (1991).

2. Illustration

We illustrate the replica method by using it to calculate the spectral distribution of the real version of the sample covariance matrix in Bai (1999, Section 2.1.2). The eigenvalue distribution of any $N \times N$ random matrix H_{jk} can be calculated by considering the generating function $Z(\mu)$ defined by

$$Z(\mu) = \int_{-\infty}^{\infty} \Big(\prod_j d\phi_j\Big) \exp\Big\{\frac{i}{2}\Big[\mu \sum_j \phi_j^2 - \sum_{jk} H_{jk}\phi_j\phi_k\Big]\Big\}, \tag{1}$$

where $\mu(x+i\varepsilon)$ implicitly contains a small positive imaginary part ε which ensures the convergence of the integrals. The integers j and k run from $1,\ldots,N$. The average normalised eigenvalue density is then given by

$$\rho(x) = \frac{2}{\pi N} \lim_{\varepsilon\to 0} \mathrm{Im} \frac{\partial}{\partial\mu}[\ln Z(\mu)]_{av}, \tag{2}$$

where $[]_{av}$ represents the average over the random variables in H_{jk}. We can connect this expression with Bai (1999) by observing that

$$\rho(x) = \frac{1}{\pi N} \lim_{\varepsilon\to 0} \mathrm{Im} \sum_{j=1}^{N} \frac{1}{\lambda_j - \mu} = \frac{1}{\pi} \lim_{\varepsilon\to 0} [\mathrm{Im} m_N(u)]_{av}, \tag{3}$$

where $m_N(\mu)$ is the Stieltjes defined in (3.1) of Bai (1999) and $\{\lambda_j, j = 1, N\}$ are the eignevalues of H_{jk}. The average in (2) is done using the replica method, which makes use of the indentity

$$[\ln Z]_{av} = \lim_{n\to 0} \frac{[Z^n]_{av} - 1}{n}. \tag{4}$$

In the right hand side of (4) the average is evaluated for integer n and then one must analytically continue to take the limit $n \to 0$. In random matrix problems this analytical continuation is straightforward, although in some physical problems, such as spin glasses, it can be more problematic. These problems occur in systems in which the phase space in the infinite system limit is partitioned so that the system is non-erogdic, see Mezard, Parisi and Virasoro (1988). This physical mechanism has no counterpart in studies of random matrices.

We will illustrate the replica method on the matrix

$$H_{jk} = \frac{1}{N} \sum_{v=1}^{p} \xi_j^v \xi_k^v, \tag{5}$$

where the real random variables $\{\xi_j^v\}$, $j = 1, \ldots, N$, $v = 1, \ldots, p$, are identically independently distributed with distribution $P(\xi_j^v)$, mean zero and variance σ^2. This matrix represents the patterns to be memorised in a neural network model of autoassociative memory, Hopfield (1982). It is also the real version of the sample covariance matrix studied in section 2.1.2 of Bai (1999). Here we have opted to study the real version because it is slightly simpler to analyse by the replica method and because the Hopfield model, which is the main application of this matrix, has real variables. To further connect with the theoretical physics literature, we have adopted the notation common within that field.

Introducing replica variables $\{\phi_{j\alpha}\}$, $j = 1, \ldots, N$ and $\alpha = 1, \ldots, n$, where n is an integer, allows us to write the average of the nth power of $Z(\mu)$ as

$$[Z^n]_{av} = \int_{-\infty}^{\infty} \Big[\prod_{j,\alpha} d\phi_{j\alpha}\Big]\Big[\prod_{j,v} P(\xi_j^v) d\xi_j^v\Big] \exp\{G\}, \tag{6}$$

where

$$G = \frac{i\mu}{2} \sum_{j,\alpha} \phi_{j\alpha}^2 - \frac{i}{2N} \sum_{\alpha,v} \Big(\sum_j \xi_j^v \phi_{j\alpha}\Big)^2. \tag{7}$$

We introduce the variables $\{x_{v\alpha}\}$, $v = 1, \ldots, p$ and $\alpha = 1, \ldots, n$ to linearise the second term in G using the Hubbard-Stratonovich transformation. This is just an integral generalisation of "completing the squares" such as

$$\exp\Big\{-\frac{i}{2N}\Big(\sum_j \xi_j^v \phi_{j\alpha}\Big)^2\Big\} = \sqrt{\frac{2}{\pi}} \int \exp\Big\{-\frac{1}{2}x_{v\alpha}^2 + \frac{1}{\sqrt{in}} x_{v\alpha} \sum_j \xi_j^v \phi_{j\alpha}\Big\} dx_{va}. \tag{8}$$

After repeatedly applying this transformation for all v and α we can intergrate over $\{\xi_j^v\}$ to obtain

$$[Z^n]_{av} = \int_{-\infty}^{\infty} \Big[\prod_{j,\alpha} d\phi_{j\alpha}\Big]\Big[\prod_{a,v} dx_{av}\Big] \exp\{\tilde{G}\}, \tag{9}$$

where

$$\tilde{G} = \frac{i\mu}{2}\sum_{j,\alpha}(\phi_{j\alpha})^2 - \frac{1}{2}\sum_{v,\alpha}(x_{va})^2 + \frac{1}{N}\sum_{j,v} f\Big(\sum_a x_{va}\phi_{j\alpha}\Big) \tag{10}$$

and $f(x) = -i\sigma^2 x^2/2$. In order to illustrate the method we assume a general form for $f(x)$ for the time being so as to represent different types of randommess. We can expand (10) for a general $f(x)$ if $y_a = x_{v\alpha}\phi_{j\alpha}$ then

$$f\Big(\sum_a y_a\Big) = f(0) + \sum_{\alpha,r} b_r y_a^r + \sum_{\alpha<\beta,r,s} b_{rs} y_a^r y_\beta^s + \sum_{\alpha<\beta<y,r,s,t} b_{rst} y_\alpha^r y_\beta^s y_\gamma^t + \cdots \tag{11}$$

without loss of generality. (In our particular case of quadratic $f(x)$, the only non-zero terms are $b_2 = -i\sigma^2/2$ and $b_{11} = -i\sigma^2$.) This alows the third term in (10) to be rewritten as

$$\begin{aligned}\frac{1}{N}\sum_{j,v} f(\sum_a x_{v\alpha}\phi_{j\alpha}) = \frac{1}{N}\sum_{j,v} f(0) &+ \frac{1}{2N}\sum_{\alpha,r} b_r\Big[\Big(\sum_v x_{v\alpha}^r + \sum_j \phi_{j\alpha}^r\Big)^2 - \Big(\sum_v x_{va}^r\Big)^2 \\ &-\Big(\sum_j \phi_{j\alpha}^r\Big)^2\Big] + \frac{1}{2N}\sum_{\alpha<\beta,r,s} b_{rs}\Big[\Big(\sum_v x_{v\alpha}^r x_{v\beta}^s + \sum_j \phi_{j\alpha}^r \phi_{j\beta}^s\Big)^2 \\ &-\Big(\sum_v x_{v\alpha}^r x_{v\beta}^s\Big)^2 - \Big(\sum_j \phi_{j\alpha}^r \phi_{j\beta}^s\Big)^2\Big] + \cdots\end{aligned} \tag{12}$$

We now introduce conjugate variables to linearise these terms, again using the Hubbard-Stratonovich transformation. The variables and their conjugates are

$$\begin{array}{cccccc} a_\alpha^{(r)} & \sum_v x_{v\alpha}^r + \sum_j \phi_{j\alpha}^r & b_\alpha^{(r)} & \sum_v x_{va}^r & c_a^{(r)} & \sum_j \phi_{j\alpha}^r \\ a_{\alpha\beta}^{(r,s)} & \sum_v x_{v\alpha}^r x_{v\beta}^s + \sum_j \phi_{j\alpha}^r \phi_{j\beta}^s, & b_{\alpha\beta}^{(r,s)} & \sum_v x_{va}^r x_{v\beta}^s \ \text{and} & c_{\alpha\beta}^{(r)} & \sum_j \phi_{j\alpha}^r \phi_{j\beta}^s. \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array} \tag{13}$$

Using these variables to linearise those in (12), then evaluating them by the method of steepest descents as p, $N \to \infty$, gives

$$\begin{array}{lll} a_\alpha^{(r)} = c\langle x_\alpha^r\rangle_2 + \langle \phi_\alpha^r\rangle_1 & b_a^{(r)} = ic\langle x_a^r\rangle_2 & c_a^{(r)} = i\langle \phi_\alpha^r\rangle_1 \\ a_{\alpha\beta}^{(r,s)} = c\langle x_\alpha^r x_\beta^s\rangle_2 + \langle \phi_a^r \phi_\beta^s\rangle_1, & b_{\alpha\beta}^{(r,s)} = ic\langle x_\alpha^r x_\beta^s\rangle_2 \ \text{and} & c_{\alpha\beta}^{(r,s)} = i\langle \phi_\alpha^r \phi_\beta^s\rangle_1 \\ \vdots \quad \vdots \quad \vdots & \vdots \quad \vdots & \vdots \quad \vdots \end{array} \tag{14}$$

where $p/N \to c \in (0,\infty)$ in the limit $p \to \infty$ and $N \to \infty$. Using the notation $D\phi = \prod_\alpha d\phi_\alpha$, the brackets $\diamondsuit_1$ and $\diamondsuit_2$ are defined by

$$\langle h\{\phi_\alpha\}\rangle_1 = \frac{\int D\phi h\{\phi_\alpha\} \exp\left\{\frac{i\mu}{2}\sum_\alpha \phi_a^2 + g_1\{\phi_a\}\right\}}{\int D\phi \exp\left\{\frac{i\mu}{2}\sum_a \phi_\alpha^2 + g_1\{\phi_\alpha\}\right\}} \tag{15}$$

and

$$\langle h\{x_\alpha\}\rangle_2 = \frac{\int D_x h\{x_\alpha\} \exp\left\{-\frac{1}{2}\sum_\alpha x_a^2 + g_2\{x_\alpha\}\right\}}{\int Dx \exp\left\{-\frac{1}{2}\sum_a x_a^2 + g\{x_a\}\right\}}, \tag{16}$$

where

$$g_1\{\phi_a\} = cf(0) + c\langle f(\sum_a \phi_a x_a)\rangle_2 \tag{17}$$

and

$$g_2\{x_a\} = \langle f(\sum_a \phi_a x_a)\rangle_1. \tag{18}$$

We can rewrite our expression for the average normalised density of states as

$$\rho(x) = \lim_{n\to 0} \frac{1}{n\pi} \mathrm{Re}\langle \sum_\alpha (\phi_\alpha)^2\rangle_1. \tag{19}$$

Using the fact that $f(x) = -i\sigma^2 x^2/2$ we can look for a self-consistent solution to (17) and (18) of the form $g_1\{\phi_\alpha\} = A\sum_\alpha(\phi_\alpha)^2$ and $g_2\{x_\alpha\} = B\sum_\alpha(x_\alpha)^2$. In this case $\rho(x)$ can be rewritten as $\rho(x) = \mathrm{Im}(A)/\pi\sigma^2$. Equations (17) and (18) can be solved self-consistently by performing the n-dimensional integrals as if n were a positive integer and then taking the limit $n \to 0$. This reveals expressions for A and B, and hence for $c > 1$,

$$\rho(x) = \begin{cases} \frac{1}{4\pi x\sigma^2}\sqrt{(b-x)(x-a)} & a \le x \le b \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

with $a = 2\sigma^2(\sqrt{c}-1)^2$ and $b = 2\sigma^2(\sqrt{c}+1)^2$. This result is of the same form as Bai (1999, equation (2.12)), if we make the changes $c \to 1/y$ and $2c\sigma^2 \to \sigma^2$. These changes are caused by different definitions of the initial random matrices, and because we are treating the real version of the matrices whereas Bai (1999) considers the complex case.

3. Summary

We have shown how the replica method can be used to calculate the eigenvalue spectrum of real random matrices. It is also possible to use this method to analyse other problems discussed in Bai (1999). For instance, in Dhesi and Jones

(1990) there is an example of how to use a perturbative scheme with the replica method to find the corrections to the spectral distribution up to $O(1/N^2)$. In Weight (1998) the replica scheme is used to analyse the properties of products of random matrices. Thus the replica technique can be viewed as a useful addition to the analytical techniques presented in Bai (1999).

Department of Mathematics and Statistics, Brunel University, Uxbridge, Middlesex, UB8 3PH U.K.
E-mail: g.j.rodgers@brunel.ac.uk

COMMENT: COMPLEMENTS AND NEW DEVELOPMENTS

Jack W. Silverstein

North Carolina State University

My good friend and colleague has done a fine job in presenting the essential tools that have been used in understanding spectral behavior of various classes of large dimensional random matrices. The Stieltjes transform is by far the most important tool. As can be seen in the paper, some limit theorems are easier to prove using them and rates of convergence of the spectral distribution can be explored using Theorem 3.1. Moreover, as will be seen below, analysis of the Stieltjes transform of the limiting spectral distribution of matrices presented in Section 2.1.3 can explain much of the distribution's properties. Also, the conjecture raised in Section 6.2 has been proven using Stieltjes transforms.

However, this is not to say the moment method can be dispensed with. Indeed, there has been no alternative way of proving the behavior of the extreme eigenvalues. This paper shows further use of moments by proving Theorem 2.10 with no restriction on $\mathbf{T}$. An attempt to prove it in Silverstein (1995) without the assumption of positive definiteness was abandoned early on in the work. Another example will be seen below concerning the preliminary work done on the rate of convergence. Moments were used. In my opinion it would be nice to develop all random matrix spectral theory without relying on moment arguments. They reveal little of the underlying behavior, and the combinatorial arguments used are frequently horrendous. Unfortunately, it appears unlikely we can remove them from our toolbox.

The remaining comments are on the matrices appearing in Theorem 2.10 when $\mathbf{T}$ is non-negative definite. Their eigenvalues are the same as those of

$$B_p \equiv \frac{1}{n} T_p^{1/2} \mathbf{X}_p \mathbf{X}_p^* T_p^{1/2},$$

(note that at this stage it is necessary to change subscripts on the matrices) where $T_p^{1/2}$ is any Hermitian square root of $\mathbf{T}_p$, and differ from those of $\mathbf{B} = \mathbf{B}_p$ in Theorem 3.4 (with $\mathbf{A} = 0$) by $|p-n|$ zero eigenvalues. When the elements of $\mathbf{X}_p$ are standardized (mean zero and $\mathbf{E}(|X_{11}|^2) = 1$), B_p is (under the assumption of zero mean) the sample covariance matrix of n samples of the p-dimensional random vector $T_p^{1/2}\mathbf{X}_{\cdot 1}$, the population matrix being of course $\mathbf{T}_p$. This represents a broad class of random vectors which includes multivariate normal, resulting in Wishart matrices. Results on the spectral behavior of B_p are relevant in situations where p is high but sample size is not large enough to ensure sample and population eigenvalues are near each other, only large enough to be on the same order of magnitude as p. The following two sections provide additional information on what is known about the eigenvalues of B_p.

1. Understanding the Limiting Distribution Through Its Stieltjes Transform

For the following, let F denote the limiting spectral distribution of B_p with Stieltjes transform $m(z)$. Then it follows that F and $\mathbf{F}$, the limiting spectral distribution of $\mathbf{B}$, satisfy

$$\mathbf{F} = (1-y)I_{[0,\infty)} + yF$$

($I_{[0,\infty)}$ denoting the indicator function on $[0,\infty)$), while $m(z)$ and $\mathbf{m}(z)$, the Stieltjes transform of $\mathbf{F}$, satisfy

$$\mathbf{m}(z) = -\frac{(1-y)}{z} + ym(z).$$

From (3.9) we find that the inverse of $\mathbf{m} = \mathbf{m}(z)$ is known:

$$z = -\frac{1}{\mathbf{m}} + y\int \frac{\tau dH(\tau)}{1+\tau\mathbf{m}},$$

and from this it can be proven (see Silverstein and Choi (1995)):

1. On $\mathbb{R}^+$, F has a continuous derivative f given by $f(x) = (1/\pi)Im\, m(x) = (1/y\pi)\lim_{z\in\mathbb{C}^+\to x} Im\,\mathbf{m}(z)$ ($\mathbb{C}^+$ denoting the upper complex plane). The density $f(x)$ is analytic wherever it is positive, and for these x, $y\pi f(x)$ is the imaginary part of the unique $\mathbf{m} \in \mathbb{C}^+$ satisfying
$$x = -\frac{1}{\mathbf{m}} + y\int \frac{\tau dH(\tau)}{1+\tau\mathbf{m}}. \qquad (1)$$
2. Intervals outside the support of f are those on the vertical axis on the graph of (1), for $m \in \mathbb{R}$, corresponding to intervals where the graph is increasing (originally observed in Marčenko and Pastur (1967)). Thus, the graph of f can be obtained by first identifying intervals outside the support, and then applying Newton's method to (1) for values of x inside the support.

3. Let $a > 0$ be a boundary point in the support of f. If a is a relative extreme value of (1) (which is always the case whenever H is discrete), then near a and in the support of f, $f \sim \sqrt{|x-a|}$. More precisely, there exists a $C > 0$ such that

$$\lim_{x \in supp\, f \to a} \frac{f(x)}{C\sqrt{|x-a|}} = \lim_{x \in supp\, f \to a} \frac{\frac{d}{dx} f(x)}{\frac{d}{dx} C\sqrt{|x-a|}} = 1$$

4. y and F uniquely determine H.
5. $F \xrightarrow{\mathcal{D}} H$ as $y \to 0$, which complements the a.s. convergence of B_p to T_p for fixed p as $n \to \infty$.
6. If $0 < b_1 < b_2$ are boundary points of the support of H with $b_1 - \epsilon$, $b_2 + \epsilon$ outside the support of H for small $\epsilon > 0$, then for all y sufficiently small there exist corresponding boundary points $a_1(y)$, $a_2(y)$ of F such that $F\{[a_1(y), a_2(y)]\} = H\{[b_1, b_2]\}$ and $[a_1(y), a_2(y)] \to [b_1, b_2]$ as $y \to 0$.

Thus from the above properties relevant information on the spectrum of $\mathbf{T}_p$ for p large can be obtained from the eigenvalues of B_p with a sample size on the same order of magnitude as p. For the detection problem in Section 5.3 the properties tell us that for a large enough sample we should be able to estimate (at the very least) the proportion of targets in relation to the number of sensors. Finding the exact number of "signal" eigenvalues separated from the $p-q$ "noise" ones in our simulations, with the gap close to the gap we would expect from F, came as a delightful suprise (Silverstein and Combettes (1992)).

2. Separation of Eigenvalues

Verifying mathematically the observed phenomenon of exact separation of eigenvalues has been achieved by Zhidong Bai and myself. The proof is broken down into two steps. The first step is to prove that, almost surely, no eigenvalues lie in any interval that is outside the support of the limiting distribution for all p large (Bai and Silverstein (1998)). Define F^A to be the empirical distribution function of the eigenvalues of the matrix A, assumed to be Hermitian. Let $H_p = F^{\mathbf{T}_n}$, $y_p = p/n$, and F^{y_p, H_p} be the limiting spectral distribution of B_p with y, H replaced by y_p and H_p. We assume the entries of $\mathbf{X}_p$ have mean zero and finite fourth moment (which are necessary, considering the results in Section 2.2.2 on extreme eigenvalues) and the matrices $\mathbf{T}_p$ are bounded for all p in spectral norm. We have then

Theorem. (Theorem 1.1 of Bai and Silverstein (1998)) *For any interval* $[a, b]$ *with* $a > 0$ *which lies in an open interval outside the support of* $F (= F^{y,H})$ *and* F^{y_p, H_p} *for all large* p *we have*

$$\text{P(no eigenvalue of } B_p \text{ appears in } [a, b] \text{ for all large } p \text{)} = 1.$$

Note that the phrase "in an open interval" was inadvertently left out of the original paper.

The proof looks closely at properties of the Stieltjes transform of F^{B_p}, and uses moment bounds on both random quadratic forms (similar to Lemma A.4 of Bai (1997)) and martingale difference sequences.

The second step is to show the correct number of eigenvalues in each portion of the limiting support. This is achieved by appealing to the continuous dependence of the eigenvalues on their matrices. Let B_p^n denote the dependence of the matrix on n. Using the fact that the smallest and largest eigenvalues of $\frac{1}{n}\mathbf{X}_p\mathbf{X}_p^*$ are near $(1-\sqrt{p/n})^2$ and $(1+\sqrt{p/n})^2$ respectively, the eigenvalues of T_p and $B_p^M n$ are near each other for suitably large M. It is then a matter of showing eigenvalues do not cross over from one support region to another as the number of samples increases from n to Mn. This work is presently in preparation.

This work should be viewed as an extension of the results in Section 2.2.2 on the extreme eigenvalues of $S_p = (1/n)X_pX_p^*$. In particular, it handles the extreme eigenvalues of B_p (see the corollary to Theorem 1.1 in Bai and Silverstein (1998)). At the same time it should be noted that the proof of exact separation relies heavily on a.s. convergence of the extreme eigenvalues of S_p. As mentioned earlier, the moment method seems to be the only way in proving Theorem 2.15. On the other hand, the Stieltjes transform appears essential in proving exact separation, partly from what it reveals about the limiting distribution.

3. Results and Conjectures on the Rate of Convergence

I will finish up with my views on the rate of convergence issue concerning the spectral distribution of sample covariance matrices raised in Section 3.2.2. The natural question to ask is: what is the speed of convergence of $W_p \equiv F^{B_p} - F^{y_p,H_p}$ to 0? Here is some evidence the rate may be $1/p$ in the case $H_p = I_{[0,\infty)}$, that is, when $B_p = S_p = (1/n)XX^*$ (Section 2.1.2).

In Jonsson (1982) it is shown that the distribution of

$$\left\{n\int x^r d(F^{S_p}(x) - \mathbf{E}(F^{S_p}(x)))\right\}_{r=1}^{\infty}$$

converges ($\mathbb{R}^\infty$) to that of a multivariate normal, suggesting an error rate of $1/p$. Continuing further, with the aid of moment analysis, the following has been observed.

Let $Y_p(x) = p\int_0^x [F^{S_p}(t) - (\mathbf{E}(F^{S_p}(t))]\,dt$. It appears that, as $p\to\infty$, $p(\mathbf{E}(F^{S_p}(x)) - F^{y_p,1_{[1,\infty)}}(x))$ converges to certain continuous function on $[0,(1+\sqrt{c})^2]$, and the covariance function $C_{Y_pY_p}(x_1,x_2) \equiv \mathbf{E}(Y_p(x_1)Y_p(x_2)) \to C_{YY}(x_1, x_2)$, continuous on $[0,(1+\sqrt{y})^2]\times[0,(1+\sqrt{y})^2]$. Both functions depend on y and $\mathbf{E}(X_{11}^4)$. Moreover, it can be verified that C_{YY} is the covariance function of a continuous mean zero Gaussian process on $[0,(1+\sqrt{c})^2]$. The uniqueness of any weakly convergent subsequence of $\{Y_p\}$ follows by the above result in Jonsson (1982) and the a.s. convergence of the largest eigenvalue of S_p (see Theorem 3.1 of Silverstein (1990)). Thus, if tightness can be proven, weak convergence of Y_p would follow, establishing the rate of convergence of 1/p for the partial sums of the eigenvalues of S_p. It should be noted that the conjecture on Y_p is substantiated by extensive simulations.

It seems that the integral making up Y_p is necessary because $\frac{\partial^2 C_{YY}}{\partial x_1 \partial x_2}(x_1,x_2)$, which would be the covariance function of $p(F^{S_p}(x)-[\mathbf{E}(F^{S_p}(x)])$ in the limit, turns out to be unbounded at $x_1=x_2$. As an illustration, when $\mathbf{E}(X_{11}^4)=3$ (as in the Gaussian case) $\frac{\partial^2 C_{YY}}{\partial x_1 \partial x_2}(x_1,x_2)=$

$$\frac{1}{2\pi^2}\ln\left[\frac{4y-((x_1-(1+y))(x_2-(1+y))+\sqrt{(4y-(x_1-(1+y))^2)(4y-(x_2-(1+y))^2)}}{4y-((x_1-(1+y))(x_2-(1+y))-\sqrt{(4y-(x_1-(1+y))^2)(4y-(x_2-(1+y))^2)}}\right]$$

for $(x_1,x_2)\in[(1-\sqrt{y})^2,(1+\sqrt{y})^2]\times[(1-\sqrt{y})^2,(1+\sqrt{y})^2]$, 0, otherwise. It therefore appears unlikely pW_p converges weakly.

Of course weak convergence of Y_p does not immediately imply $\alpha(p)W_p\to 0$ for $\alpha(p)=o(p)$. It only lends support to the conjecture that $1/p$ is the correct rate. Further work is definitely needed in this area.

Acknowledgement

This work is supported by NSF Grant DNS-9703591.

Department of Mathematics, North Carolina State University, Raleigh, NC, U.S.A.

REJOINDER

Z. D. Bai

Thanks to Professor Jack Silverstein and Dr. G. J. Rodgers for their additions to developments in the theory of spectral analysis of large dimensional random matrices not reported on in my review paper. I would like to make some remarks on the problems arising from their comments.

1. Spectrum Separation of Large Sample Covariance Matrices

Jack Silverstein reported a new result on spectrum separation of large sample covariance matrices obtained in Bai and Silverstein (1998), after my review paper was written. It is proved there that under very general conditions, for any closed interval outside the support of the limiting spectral distribution of a sequence of large dimensional sample covariance matrices, and with probability 1 for all large n, the sample covariance matrix has no eigenvalues falling in this interval. He also reported that a harder problem of exact spectrum separation is under our joint investigation. Now, I take this opportunity to report that this problem has been solved in Bai and Silverstein (1999). More specifically, the exact spectrum separation is established under the same conditions of Theorem 1.1 of Bai and Silverstein (1998).

1.1. Spectrum separation of large sample covariance matrices

Our setup and basic assumptions are the following.

(a) X_{ij}, $i, j = 1, 2, ...$ are independent and identically distributed (i.i.d.) complex random variables with mean 0, variance 1 and finite 4th moment;
(b) $n = n(p)$ with $y_n = p/n \to y > 0$ as $n \to \infty$;
(c) For each n, $\mathbf{T}_n$ is a $p \times p$ Hermitian nonnegative definite matrix satisfying $H_n \equiv F^{\mathbf{T}_n} \xrightarrow{\mathcal{D}} H$, a cumulative distribution function (c.d.f.);
(d) $\|\mathbf{T}_n\|$, the spectral norm of $\mathbf{T}_n$, is bounded in n;
(e) $\mathbf{S}_n = n^{-1}\mathbf{T}_n^{1/2}\mathbf{X}_n\mathbf{X}_n^*\mathbf{T}_n^{1/2}$, $\underline{\mathbf{S}}_n = n^{-1}\mathbf{X}_n^*\mathbf{T}_n\mathbf{X}_n$, where $\mathbf{X}_n = (X_{ij})$, $i = 1, \ldots, p$, $j = 1, \ldots, n$, and $\mathbf{T}_n^{1/2}$ is a Hermitian square root of T_n.

The matrix $\mathbf{S}_n$ is of major interest and the introduction of the matrix $\underline{\mathbf{S}}_n$ is for mathematical convenience. Note that

$$F^{\underline{\mathbf{S}}_n} = \Big(1 - y_n\Big) I_{[0,\infty)} + y_n F^{\mathbf{S}_n}$$

and

$$m_{F^{\underline{\mathbf{S}}_n}}(z) = -\frac{1 - y_n}{z} + y_n m_{F^{\mathbf{S}_n}}(z).$$

As previously mentioned, under conditions (a) - (e), the limiting spectral distribution (LSD) of $\mathbf{S}_n$ exists and the Stieltjes transform of the LSD of $\underline{\mathbf{S}}_n$ is the unique solution, with nonnegative imaginary part for z on the upper half plane, to the equation

$$z_{y,H}(m) = -\frac{1}{m} + y \int \frac{t}{1 + tm} dH(t).$$

The LSD of $\underline{\mathbf{S}}_n$ is denoted by $F^{y,H}$. Then, for each fixed n, F^{y_n,H_n} can be regarded as the LSD of a sequence of sample covariance matrices for which the LSD of the population covariance matrices is H_n and limit ratio of dimension to sample size is y_n. Its Stieltjes transform is then the unique solution with nonnegative imaginary part, for z on the upper half plane, to the equation

$$z_{y_n,H_n}(m) = \frac{1}{m}\left(-1 + y_n \int \left(\frac{tm}{1 + tm}\right) dH_n(t)\right) \tag{1}$$

$$z'_{y_n,H_n}(m) = \frac{1}{m^2}\left(1 - y_n \int \left(\frac{tm}{1 + tm}\right)^2 dH_n(t)\right). \tag{2}$$

It is easy to see that for any real $x \neq 0$, the function $m_{F^{y_n,H_n}}(x)$ and its derivative are well defined and continuous provided $-1/x$ is not a support point of H_n.

Under the further assumption that

(f) the interval $[a, b]$ with $a > 0$ lies in an open interval outside the support of F^{y_n,H_n} for all large n,

Bai and Silverstein (1998) proved that with probability one, for all large n, $\mathbf{S}_n$ has no eigenvalues falling in $[a, b]$.

To understand the meaning of exact separation, we give the following description.

1.2. Description of exact separation

From (1), it can be seen that F^{y_n,H_n} and its support tend to $F^{y,H}$ and the support of it, respectively. We use F^{y_n,H_n} to define the concept exact separation in the following.

Denote the eigenvalues of $\mathbf{T}_n$ by $0 = \lambda_1(\mathbf{T}_n) = \cdots = \lambda_h(\mathbf{T}_n) < \lambda_{h+1}(\mathbf{T}_n) \leq \cdots \leq \lambda_p(\mathbf{T}_n)$ ($h = 0$ if $\mathbf{T}_n$ has no zero eigenvalues). Applying Silverstein and Choi (1995), the following conclusions can be made. From (1) and (2), one can see that $mz_{y_n,H_n}(m) \to -1+y_n(1-H_n(0))$ as $m \to -\infty$, and $mz_{y_n,H_n}(m) > -1+y_n(1-H_n(0))$ for all $m < -M$ for some large M. Therefore, when m increases along the real axis from $-\infty$ to $-1/\lambda_{h+1}(\mathbf{T}_n)$, the function $z_{y_n,H_n}(m)$ increases from 0 to a maximum and then decreases to $-\infty$ if $-1 + y_n(1 - H_n(0)) \geq 0$; it decreases directly to $-\infty$ if $-1 + y_n(1 - H_n(0)) < 0$, where $H_n(0) = h/p$. In the latter case, we say that the maximum value of z_{y_n,H_n} in the interval $(-\infty, -\lambda_{h+1}(\mathbf{T}))$ is 0. Then, for $h < k < p$, when m increases from $-1/\lambda_k(\mathbf{T}_n)$ to $-1/\lambda_{k+1}(\mathbf{T}_n)$, the function z_{y_n,H_n} in (1) either decreases from ∞ to $-\infty$, or decreases from ∞ to a local minimum, then increases to a local maximum and finally decreases to $-\infty$. Once the latter case happens, the open interval of z_{y_n,H_n} values from the minimum to the maximum is outside the support of F^{y_n,H_n}. When m increases from $-1/\lambda_p(\mathbf{T}_n)$ to 0, the z value decreases from ∞ to a local minimum and then increases to ∞. This local minimum value determines the largest boundary of the support of F^{y_n,H_n}. Furthermore, when m increases from 0 to ∞, the function $z_{y_n,H_n}(m)$ increases from $-\infty$ to a local maximum and then decreases to 0 if $-1 + y_n(1 - H_n(0)) > 0$; it increases directly from $-\infty$ to 0 if $-1 + y_n(1 - H_n(0)) \leq 0$. In the latter case, we say that the local maximum value of z_{y_n,H_n} in the interval $(0, \infty)$ is 0. The maximum value of z_{y_n,H_n} in $(-\infty, -\lambda_{h+1}(\mathbf{T})) \cup (0, \infty)$ is the lower bound of the support of F^{y_n,H_n}.

Case 1. $y(1-H(0)) > 1$. For all large n, we can prove that the support of $F^{y,H}$ has a positive lower bound x_0 and $y_n(1 - H_n(0)) > 1$, $p > n$. In this case, we can prove that $\mathbf{S}_n$ has $p - n$ zero eigenvalues and the nth largest eigenvalues of $\mathbf{S}_n$ tend to x_0.

Case 2. $y(1-H(0)) \leq 1$ or $y(1-H(0)) > 1$ but $[a,b]$ is not in $[0,x_0]$. For large n, let $i_n \geq 0$ be the integer such that

$$\lambda_{i_n}^{T_n} > -1/m_{F^{y,H}}(b) \qquad \text{and} \qquad \lambda_{i_n+1}^{T_n} < -1/m_{F^{y,H}}(a).$$

It is seen that only when $m_{F^{y,H}}(b) < 0$, the exact separation occurs. In this case, we prove that

$$P(\lambda_{i_n}^{\mathbf{S}_n} > b \quad \text{and} \quad \lambda_{i_n+1}^{\mathbf{S}_n} < a \quad \text{for all large } n\ ) = 1.$$

This shows that with probability 1, when n is large, the number of eigenvalues of $\mathbf{S}_n$ which are greater than b is exactly the same as that of the eigenvalues of $\mathbf{T}_n$ which are greater than $-1/m_{F^{y,H}}(b)$, and contrarily, the number of eigenvalues of $\mathbf{S}_n$ which are smaller than a is exactly the same as that of the eigenvalues of $\mathbf{T}_n$ which are smaller than $-1/m_{F^{y,H}}(a)$.

1.3. Strategy of the proof of exact spectrum separation

Consider a number of sequences of sample covariance matrices of the form

$$\mathbf{S}_{n,k} = (n+kM)^{-1}\mathbf{T}_n^{1/2}\mathbf{X}_{n,k}\mathbf{X}_{n,k}^*\mathbf{T}_n^{1/2},$$

where $M = M_n$ is an integer such that $M/n \to c > 0$, for some small $c > 0$, and $\mathbf{X}_{n,k} = (X_{ij})$ with dimension $p \times (n+kM)$.

We need to prove the following.

(i) Define $y_k = y/(1+kc)$ and $a_k < b_k$ by

$$m_{F^{y,H}}(a) = m_{F^{y_k,H}}(a_k) \quad \text{and} \quad m_{F^{y,H}}(b) = m_{F^{y_k,H}}(b_k).$$

We show that when $c > 0$ is small enough,

$$P(\lambda_{\ell_n}(\mathbf{S}_{n,k}) < a_k \quad \text{and} \quad \lambda_{\ell_n+1}(\mathbf{S}_{n,k}) > b_k \quad \text{for all large } n) = 1,$$

and thus that

$$P(\lambda_{\ell_n}(\mathbf{S}_{n,k+1}) < a_{k+1} \quad \text{and} \quad \lambda_{\ell_n+1}(\mathbf{S}_{n,k+1}) > b_{k+1} \quad \text{for all large } n) = 1.$$

(ii) Let K be so large that $p/(n+KM) := y_{Kn} \to y_K = y/(1+Kc) < (-1/m_{F^{y,H}}(b) + 1/m_{F^{y,H}}(a)/(9\max_n \|\mathbf{T}_n\|)$. Then, by Corollary 7.3.8 of Horn and Johnson (1985), we have

$$\begin{aligned}
&\max_{i\leq p} |\lambda_i(\mathbf{S}_{n,K}) - \lambda_i(\mathbf{T}_n)| \leq \|T_n^{1/2}(\mathbf{S}_{n,K} - \mathbf{I})\mathbf{T}_n^{1/2}\| \\
&\leq \max_n \|\mathbf{T}_n\| \max(\lambda_p(\mathbf{S}_{n,K}) - 1, 1 - \lambda_1(\mathbf{S}_{n,K})) \\
&\to \max_n \|\mathbf{T}_n\|(2y_K + y_K^2) < (-1/m_{F^{y,H}}(b) + 1/m_{F^{y,H}}(a)/3. \qquad (3)
\end{aligned}$$

From (ii), it follows that with probability 1, for all large n, $\lambda_{i_n+1}(\mathbf{S}_{n,K}) > (b_K + a_K)/2$ and $\lambda_{i_n}(\mathbf{S}_{n,K}) < (b_K + a_K)/2$. Then by Bai and Silverstein (1998), $\lambda_{i_n+1}(\mathbf{S}_{n,K}) > b_K$ and $\lambda_{i_n}(\mathbf{S}_{n,K}) < a_K$. That is, the exact spectrum separation holds for the sequence $\{\mathbf{S}_{n,K}, n = 1, \ldots$. Applying (i), the exact spectrum separation remains true for any sequence $\{\mathbf{S}_{n,k}, n = 1, \ldots\}$

2. On Replica Method

People working in the area of spectral analysis of large dimensional random matrices are aware that the theory was motivated by early findings, laws or conjectures, in theoretical physics, see the first paragraph of the introduction of my review paper (BaiP, hereafter). However, very few papers in pure probability or statistics refer to later developments in theoretical physics. Therefore, I greatly appreciate the relation of later developments in theoretical physics by G. J. Rodgers in his comments (RodC, hereafter), including the replica method and some valuable references.

From my point of view, the replica method starts at the same point as does the method of Stieltjes transform, analyzes with different approaches, and finds the same conclusions. At first, we note that the function $Z(\mu)$ defined in (1) of RodC is in fact $(2\pi i)^{N/2}\det^{1/2}(\mathbf{H} - \mu\mathbf{I})$. From this, one can derive that

$$\frac{2}{\pi N}\frac{\partial}{\partial\mu}\log Z(\mu) = \frac{1}{\pi N}\sum_{j=1}^{N}\frac{1}{\lambda_j - \mu} = \pi^{-1}m_n(\mu),$$

where $m_n(\cdot)$ is defined in (3.3) of BaiP. Note that $[Z^n(\mu)]_{av} = EZ^n(\mu)$. Consequently, the function in (2) of RodC is in fact

$$\rho(\mu) = \mathrm{Im}\frac{2}{\pi N}\frac{\partial}{\partial\mu}\log EZ^n(\mu).$$

For all large N, we should have $\rho(\mu) \sim \pi^{-1}\mathrm{Im}Em_n(\mu)$, which is asymptotically independent of n. This shows that the two methods start from the same point.

The method of Stieltjes transformation analyzes the resolvent of the random matrices by splitting

$$m_n(\mu) = \frac{1}{N}tr(\mathbf{H} - \mu\mathbf{I})^{-1}$$

into a sum of weakly dependent terms, while the replica method continues its analysis on the expected function $[Z^n(\mu)]_{av}$.

Now, we consider the Hubbard-Stratonovich transformation, in which a set of i.i.d standard normal variables $x_{\alpha j}$ are used to substitute for the variables

$$\sigma^{-1}\sum_{i=1}^{N}\xi_i^j\phi_{i\alpha}\left(\sum_{i=1}^{N}\phi_{i\alpha}^2\right)^{-1/2}.$$

The validity of this normal approximation is a key point in the replica method and might be the reason to call it "non-rigorous" in RodC.

For each fixed α and j, it is not difficult to show that as $N \to \infty$, the variable $\sigma^{-1}\sum_{i=1}^{N}\xi_i^j\phi_{i\alpha}$ is asymptotically normal for $\phi_{i,\alpha}$'s satisfying $\sum_{i=1}^{N}\phi_{i\alpha}^2 = 1$, except in a small portion on the unit sphere. However, I do not know how to show the asymptotic independence between $\sigma^{-1}\sum_{i=1}^{N}\xi_i^j\phi_{i\alpha}$ for different (j, α)'s. If this can be done, then many problems in the spectral analysis of large dimensional random matrices, say, the circular law under the only condition of the finite second moment, can be reduced to the normal case, under which the problems are well-known or easier to deal with. More specifically, the conjectures are the following.

Conjecture 1. Let $\mathbf{X}$ be an $n \times N$ matrix with i.i.d. entries of mean zero and variance 1, and let $\mathbf{H}$ be uniform distributed on the $p \times n$ $(p \le n)$ matrix space of p orthonormal rows. Then as p, n, N proportionally tend to infinity, the $p \times N$ entries of $\mathbf{HX}$ are *asymptotically* **i.i.d.** normal.

Of course, there is a problem on how to define the terminology **asymptotically i.i.d.** since the number of variables goes to infinity. For use in spectral analysis of large dimensional random matrices, we restate Conjecture 1 as the following.

Conjecture 2. Let $\mathbf{X}$ be an $n \times N$ matrix with i.i.d. entries of mean zero and variance 1, and let $\mathbf{E}$ be uniform distributed on the $n \times n$ orthogonal matrix space. Then as n, N proportionally tend to infinity, the limiting behavior of all spectrum functionals of the matrix $\mathbf{HX}$ are the same as if all entries of $\mathbf{X}$ are i.i.d. normal.

More specifically, we have

Conjecture 3. Let $\mathbf{X}$ be an $n \times N$ matrix with i.i.d. entries of mean zero and variance 1. There exists an $n \times n$ orthogonal matrix $\mathbf{H}$ such that as n, N proportionally tend to infinity, the limiting behavior of all spectrum functionals of the matrix $\mathbf{HX}$ are the same as if all entries of $\mathbf{X}$ are i.i.d. normal.

This seems to be a very hard but interesting problem.

Additional References

Bai, Z. D. (1997). Circular Law. *Ann. Probab.* **25**, 494-529.

Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statist. Sinica*, previous paper.

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

Bai, Z. D. and Silverstein, J. W. (1999). Exact separation of eigenvalues of large dimensional sample covariance matrices. Accepted by *Ann. Probab.*

Bray, A. J. and Rodgers, G. J. (1988). Diffusion in a sparsely connected space: a model for glassy relaxation. *Phys. Rev. B* **38**, 11461-11470.

Dhesi, G. S. and Jones, R. C. (1990). Asymptotic corrections to the Wigner semicircular eigenvalue spectrum of a large real symmetric random matrix using the replica method. *J. Phys. A* **23**, 5577-5599.

Edwards, S. F. (1970). Statistical mechanics of polymerized materials. *Proc. 4th Int. Conf. on Amorphous Materials* (Edited by R. W. Douglas and B. Ellis), 279-300. Wiley, New York.

Edwards, S. F. and Jones, R. C. (1976). Eigenvalue spectrum of a large symmetric random matrix. *J. Phys. A* **9**, 1595-1603.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Ac. Sci. USA* **79**, 2554-2558.

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

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

Mezard, M., Parisi, G. and Virasoro, M. (1988). *Spin Glass Theory and Beyond*. World Scientific, Singapore.

Mirlin, A. D. and Fyodorov, Y. V. (1991). Universality of level correlation function of sparse random matrices. *J. Phys. A* **24**, 2273-2286.

Rodgers, G. J. and Bray, A. J. (1988). Density of states of a sparse random matrix *Phys. Rev. B* **37**, 3557-3562.

Rodgers, G. J. and De Dominicis, C. (1990). Density of states of sparse random matrices. *J. Phys. A* **23**, 1567-1573.

Silverstein, J. W. (1995). Strong convergence of the eimpirical distribution of eigenvalues of large dimensional random matrices *J. Multivariate Anal.* **5**, 331-339.

Silverstein, J. W. (1990). Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *Ann. Probab.* **18**, 1174-1194.

Silverstein, J. W. and Choi, S. I. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.* **54**, 295-309.

Silverstein, J. W. and Combettes, P. L. (1992). Signal detection via spectral theory of large dimensional random matrices. *IEEE Trans. Signal Processing* **40**, 2100-2105.

Sommers, H. J., Crisanti, A., Sompolinsky, H. and Stein, Y. (1988). Specturm of large random asymmetric matrices. *Phys. Rev. Lett.* **60**, 1895-1898.

Weight, M. (1998). A replica approach to products of random matrices. *J. Phys. A* **31**, 951-961.