Journal club - Random matrix theory in statistics: A review

Geneviève Robin

March 12, 2017

The article "Random matrix theory in statistics: A review" was written by D. Paul and A. Aue and published in the Journal of Statistical Planning and Inference in 2015. Random Matrix Theory (RMT) is interested among other topics in describing the asymptotic behavior of the singular values and singular vectors of random matrices. Random matrices emerge in many statistical problems, that can be treated using results from random matrix theory. Applications include hypothesis testing, covariance estimation and dimensionality reduction. I present the most classical results reviewed by Paul & Aue and their application two three statistical problems. I focus on real-valued random matrices, but most of the results presented here hold for complex-valued matrices and their extensions can be found in the original paper.

1 Random matrices in statistics

The study of the spectrum and eigenvectors of random matrices is found in PCA and MANOVA.

1.1 Principal Component Analysis (PCA)

Consider $X \in \mathbb{R}^p$ a random vector with covariance matrix $\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$. The goal of PCA is to learn a low dimensional representation of the vector X such that the residual variance is minimized. This is solved by finding a sequence of orthonormal vectors v_k, $k = 1, \dots, p$ such that $v_k = \text{argmax}\, \text{Var}(v_k X^T)$, $\|v_k\| = 1$ and $v_k v_j^T = 0$ for all $j = 1, \dots, k-1$. The v_k are equivalently defined as a sequence of eigenvectors of Σ. In practice we observe n realizations $X^{(i)}$, $i = 1, \dots, n$ of X and Σ is unknown. We estimate the v_k by their empirical counterparts $\hat{v}_k$ defined as a sequence of orthonormal eigenvectors of $S = n^{-1} \sum_{i=1}^n (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T$, with eigenvalues $\hat{l}_k$. S is a random matrix, and when X is multivariate normal, nS follows a Wishart distribution $W_p(n-1, \Sigma)$.

1.2 Multivariate Analysis of Variance (MANOVA)

MANOVA is the multivariate extension of ANOVA for testing the equality of the means of multivariate random vectors. Assume the existence of g normal populations $X^1, \dots, X^g$ with means $\mu_1, \dots, \mu_g$ and common covariance matrix Σ, and that we observe samples of size $n_1, \dots, n_g$ of these populations. To test the null hypothesis H_0:$\mu_1 = \dots = \mu_g$ against the alternative H_1 that there exist $i, j \in \{1, \dots, g\}$ such that $\mu_i \neq \mu_j$ we define the so-called F matrix (multivariate counterpart

of the F statistics) defined as UV^{-1}, where $U = \sum_{i=1}^{g} n_i(\bar{X}^i - \bar{X})(\bar{X}^i - \bar{X})^T$ is the between groups covariance matrix and $V = \sum_{i=1}^{g} sum_{j=1}^{n_i}(X_j^i - \bar{X}^i)(X_j^i - \bar{X}^i)^T$ is the within groups covariance matrix. Under the null hypothesis, U and V are independent Wishart, and UV^{-1} is referred to as the double Wishart matrix. Let $l_{\max}$ be the largest eigenvalue of UV^{-1}. A classical test is to reject H_0 for large values of $l_{\max}$. If we know the distribution of $l_{\max}$ under H_0 we can derive an exact test.

2 Asymptotics results from RMT

In these two applications it is of interest to describe the behavior of the spectrum and eigenvectors of random matrices. Asymptotic results for the spectrum of random matrices have been proven in RMT. The asymptotics is defined for a matrix $X \in \mathbb{R}^{p\times n}$ as the sample size $n \to \infty$ and the number of variables $p(n) \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

2.1 Empirical Spectral Distribution (ESD)

Consider a random matrix $X \in \mathbb{R}^N$ and denote by $l_1, \ldots, l_N$ its eigenvalues. The ESD of X is the fuction defined as $1/N \sum_{i=1}^{N} \delta_i$, where δ_y is the Dirac mass at y. If X is Hermitian, the l_i are in $\mathbb{R}$ and we define the Empirical Distribution Function (EDF) of X $F^X(x) = 1/N \sum_{i=1}^{N} \mathbb{1}_{\{l_i \leq x\}}$. Many statistics can be written as linear spectral statistics in the form of $\int g(x) dF^X(x)$, where g is a sufficiently regular function. For example $\log\det(X) = N \int \log(x) dF^X(x)$ and $\text{trace}(X^k) = N \int x^k dF^X(x)$, for $k \geq 0$. RMT is in particular interested in describing the asymptotic convergence of the ESD to a proper probality distribution. In this section I state some classical convergence results for special classes of random matrices.

2.2 Wigner matrices

Wigner matrices were introduced by Wigner in 1955 to describe the spectra of heavy atoms in Quantum Mechanics. They are defined as follows. $X \in \mathbb{C}^{N\times N}$ or $\mathbb{R}^{N\times N}$ Hermitian or symmetric with independent diagonal entries X_{ii} and independent off-diagonal entries X_{ij}, $i < j$ with mean 0 and variance 1. The simplest example of Wigner matrices is the Gaussian Unitary Ensemble defined by $X_{ii} \sim \mathcal{N}(0, 1)$ and $X_{ij} \sim \prime, \infty/\in\mathcal{I}_\in$ for $i < j$ (the X_{ij} are complex). In this ensemble the finite sample joint distribution of the eigenvalues of X is explicit and given by $f(l_1, \ldots, l_N) = C_N \prod_{1\leq j<k\leq N} |l_j - l_k|^2 \exp\left(-1/2 \sum_{i=1}^{N} l_i\right)$.

Theorem 1 *Wigner 1958 Let $X \in \mathbb{C}^{N\times N}$ be a Wigner matrix. Assume X has i.i.d. real valued diagonal entries with $\mathbb{E}[X_{ii}] = 0$ and $\text{Var}[X_{ii}] = 1$. Assume further that X has i.i.d off-diagonal entries such that $\mathbb{E}[X_{ii}] = 0$ and $\text{Var}[X_{ii}] = 1$ for $i < j$.*
Then, as $N \to +\infty$, the ESD of $1/\sqrt{N}X$ $\mu_{X/\sqrt{N}}$ converges in distribution to the semi-circle law defined as $f_{sc}(x) = 1/(2\pi)\sqrt{4 - x^2}\mathbb{1}_{[-2,2]}(x)$.
Moreover if $\mathbb{E}|X_{ij}|^4 < +\infty$ then the largest eigenvalue $l_{max}/\sqrt{N} \overset{a.s.}{\to} 2$ and the smallest eigenvalue $l_{min}/\sqrt{N} \overset{a.s.}{\to} -2$.

This result is universal in the sense that it does not depend on the distribution of the entries of X. On the convergence rate of the ESD to the limiting semi-circle law depends on the distribution of

the X_{ij}. A first famous proof strategy is the method of moments where one shows that the moments of the ESD $\beta_{k,N}(\mu_{X/\sqrt{N}}) = \int x^k dF^{X/\sqrt{N}}(x) = \text{trace}(X^k)/N^{1+p/2}$ converge to the moments of the semi-circle law for all k. Another method consists in showing that the Stieltjes transform of $\mu_{X/\sqrt{N}}$ defined as $S_\mu(z) = \int 1/(\mu(x) - z)\mu(dx)$ converges to that of the semi-circle law. Wigner's theorem is illustrated in Figure 1, where the ESD of a gaussian Wigner matrix is represented along with the semi-circle law for increasing values of N (from left to right $3, 10, 50$).

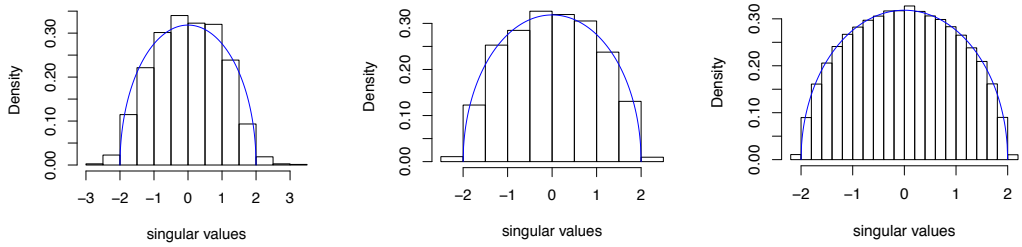

Figure 1: Empirical distribution of the bulk spectrum of a Wigner matrix of size from left to right $3, 10, 50$ and the semi-circle law.

2.3 Wishart matrices

Wishart matrices were introduced by Wishart in 1928. Consider $X \in \mathbb{C}^{p\times n}$ an observation matrix with independent columns $X_j = (X_{ij})_{i=1}^p$. $S = 1/nXX^T$ is the (uncentered) sample covariance matrix, and nS is referred to as the Wishart matrix, because when the X_i are i.i.d multivariate gaussians with mean 0 and covariance matrix Σ, S follows a Wishart distribution $W_p(n-1, \Sigma)$ with $, -1$ degreees of freedom and centered in Σ.

Theorem 2 *Marcenko-Pastur Let $X \in \mathbb{C}^{p\times n}$ be a Wigner matrix. Assume X has i.i.d. entries with $\mathbb{E}[X_{ij}] = 0$ and $\text{Var}[X_{ij}] = 1$. Assume further that $p/n \to \gamma \in (0, +\infty)$ as $n \to +\infty$.*
Then, as $n \to +\infty$, the ESD of $S = 1/nXX^*$ μ_{1/nXX^*} converges almost surely in distribution to the Marcenko-Pastur law denoted by f_γ.
If $\gamma \leq 1$ (at least as many observations as variables) then

$$f_\gamma(x) = \frac{\sqrt{(b_+(\gamma) - x)(x - b_-(\gamma))}}{2\pi\gamma x} \mathbb{1}_{[b_-(\gamma), b_+(\gamma)]}(x),$$

with $b_{+/-}(\gamma) = (1 +/- \sqrt{\gamma})^2$.
If $\gamma > 1$ (less observations than variables) then the Marcenko-Pastur law is a mixture of a Dirac mass at 0 (S in singular) and $f_{1/\gamma}$ with weights $1 - 1/\gamma$ and $1/\gamma$ respectively.

$$f_\gamma(x) = \frac{\sqrt{(b_+(\gamma) - x)(x - b_-(\gamma))}}{2\pi\gamma x} \mathbb{1}_{[b_-(\gamma), b_+(\gamma)]}(x),$$

with $b_{+/-}(\gamma) = (1 +/- \sqrt{\gamma})^2$.

Figures 2 and 3, where the ESD of a gaussian Wishart matrix is represented along with the Marcenko-Pstur law for increasing values of N (from left to right $3, 10, 50$), with $\gamma \leq 1$ in Figure 2 and $\gamma > 1$ in Figure 3, illustrate the Marcenko-Pastur theorem.

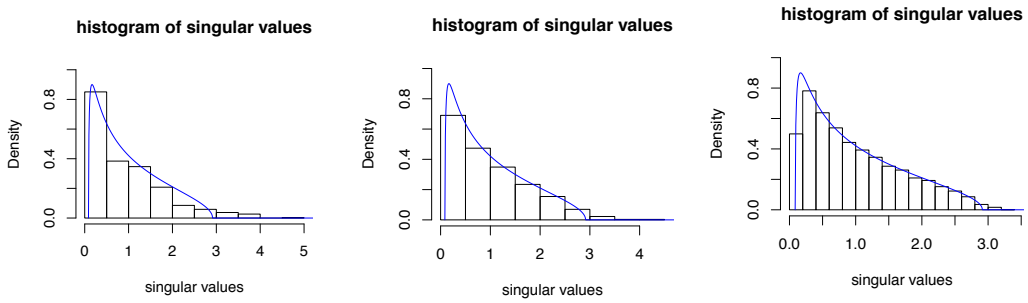

Figure 2: Empirical distribution of the bulk spectrum of a Wishart matrix of size from left to right $3, 10, 50$ and $\gamma < 1$ and the Marcenko-Pastur law.

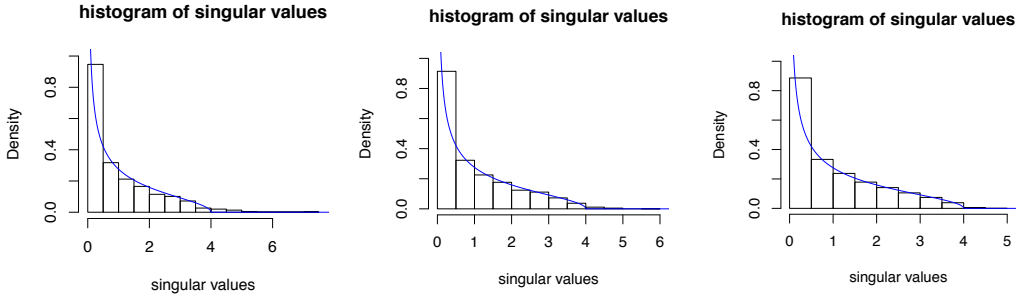

Figure 3: Empirical distribution of the bulk spectrum of a Wishart matrix of size from left to right $3, 10, 50$ and $\gamma = 1$ and the Marcenko-Pastur law.

2.4 The double Wishart problem

Consider $X_1 \in \mathbb{C}^{p \times n_1}$ and $X_2 \in \mathbb{C}^{p \times n_2}$ to independent matrices with independent columns. Assume that the entries of X_1 and X_2 have mean 0 and variance 1. Then $X_1 X_1^*$ and $X_2 X_2^*$ are independent Wishart matrices and $X_1 X_1^* (X_2 X_2^*)^{-1}$ is referred to as the double Wishart. Its eigenvalues are in one-to-one correspondence to those of $X_1 X_1^* (X_1 X_1 + X_2 X_2^*)^{-1}$. Yin et al. (1983) proved the existence of a limiting ESD for the double Wishart.

3 The edge of the spectrum

We now turn to the asymptotic behavior of the largest eigenvalue of the random matrices defined in the previous section.

3.1 The Tracy-Widom law

We first define the Tracy-Widom law which is involved in the asymptotic results. We focus on the TW_1 law relevant for real-valued random variables and defined as

$$TW_1(s) = \exp\left(-\frac{1}{2}\int_s^\infty \left(q(x) - (x-s)q^2(x)\right) dx\right),$$

for $s \in \mathbb{R}$ where q satisfies the Poincaré II equation $q''(x) = xq(x) + 2q^3(x)$ under the constraint $q(x) - A(x) \to 0$ and A is the Airy function (see Olver, 1974).
The left tail of the Tracy-Widom law decreases as a Gaussian while the right tail has a slower rate.

3.2 Asymptotic distribution of largest eigenvalue

The asymptotic distribution of the largest eigenvalue of Wishart and double Wishart matrices is given by the two following theorems.

Theorem 3 *Largest eigenvalue of Wishart matrix Let $X \in \mathbb{R}^{p\times n}$ be an observation matrix with i.i.d. entries of mean 0 and variance 1, and denote by l_1 the largest eigenvalue of XX^T. Assume further that $p/n \to \gamma \in (0,1]$ as $n \to \infty$.*
Then, as $n \to \infty$, the quantity $(l_1 - \mu_{n,p})/\sigma_{n,p}$ converges in law to the TW_1 distribution, with $\mu_{n,p} = (\sqrt{n-1} + \sqrt{p})^2$ and $\sigma_{n,p} = (\sqrt{n-1} + \sqrt{p})(1/\sqrt{n-1} + 1/\sqrt{p})^{1/3}$.

Figure 4 illustrates this theorem for X with i.i.d. standard normal Gaussian entries and $\gamma = 0.6$, representing the empirical distribution of $(l_1 - \mu_{n,p})/\sigma_{n,p}$ for increasing values of n ($10, 50, 200$ from left to right).

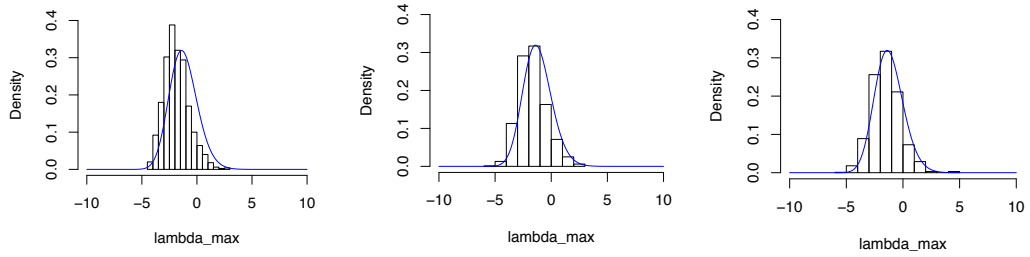

Figure 4: Empirical distribution of the largest eigenvalue of a Wishart matrix of size from left to right $10, 50, 200$ and $\gamma = 0.6$ and the Tracy-Widom law.

Theorem 4 *Largest eigenvalue of double Wishart matrix Let $X_1 \in \mathbb{R}^{n\times p}$, $X_2 \in \mathbb{R}^{n\times p}$ be two matrices. Define $U = X_1X_1^T$ and $V = X_2X_2^T$, and θ the largest eigenvalue of $U(U+V)^{-1}$. Assume that $n_2(p) \to \infty$ as $p \to \infty$ and $n_1(p) \to \infty$ as $p \to \infty$, such that* min.

3.3 Universality

A lot of effort has been focused on proving the universality of the convergence of the ESD and of the behavior of the largest eigenvalue so that results can be applied to other distributions (Poisson,

etc.). Let us give some precisions about which results depend on the distribution of the entries and which don't. The first order convergence of the ESD is universal as long as the entries are centered and scaled. On the contrary the convergence rate is not universal. The limiting distribution of the normalized largest eigenvalue is not universal either. The Tracy-Widom limit requires that the first fourth moments of the distribution match those of a Gaussian distribution. This was proven by Tao and Vu (2012) and is referred to as the "Four Moments Theorem".
As for heavy-tailed random matrices, Soshnikov (2004, 2006) showed that in the absence of a finite fourth moment, the behavior of the largest eigenvalue is determined by the behavior of the largest entry of the matrix.

4 Applications

We now give some applications of RMT in statistics.

4.1 Signal detection

Assume that we observe realizations of a random vector $X \in \mathbb{R}^p$ sampled from a multivariate distribution $\mathcal{N}(\mu, \Sigma)$. One can test the existence of a one-dimensional signal over a Gaussian white noise. We test the null hypothesis $H_0 : \Sigma = I_p$ against the aleternative $H_1 : \Sigma = l\theta\theta^T + I_p$ where $\theta \in \mathbb{R}^p$. Denote by S the sample covariance matrix of X and $\hat{l}$ the largest eigenvalue of S. Under H_0 the asymptotic distribution of $\hat{l}$ is given by the Tracy-Widom law TW_1. An asymptotic test of level $1-\varepsilon$, $0 < \varepsilon < 1$ is to accept H_0 if $\hat{l}$ is smaller or equal to the $1-\varepsilon$-quantile of TW_1 (that can be found in tables in R or Matlab), and to accept it otherwise.

4.2 Spiked model

We now complicate a little bit the model by considering a multi-dimensional signal over a Gaussian white noise: $\Sigma = \sum_{i=1}^{M} l_i\theta_i\theta_i^T + \sigma^2 I_p$. RMT results can be used to show the inconsistence of PCA in this framework.

Theorem 5 *Consider the covariance matrix* $\Sigma = \sum_{i=1}^{M} l_i\theta_i\theta_i^T + \sigma^2 I_p \mathbb{R}^{p\times p}$, $l_1 \leq \ldots \leq l_M > 1$. *Let* $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^p$ *be an n-sample from the multivariate Gaussian distribution* $\mathcal{N}(0, \Sigma)$. *Define*

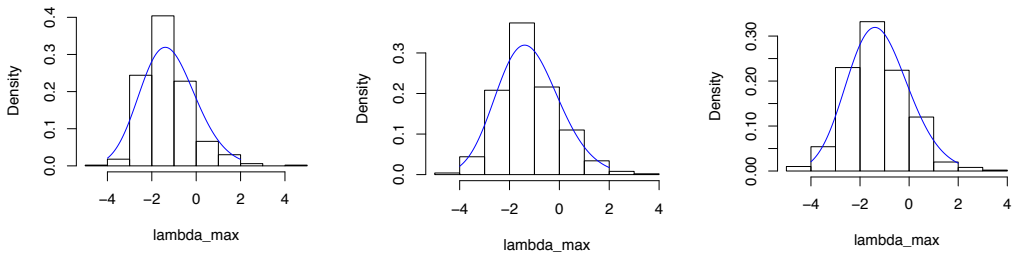

Figure 5: Empirical distribution of the largest eigenvalue of a double Wishart matrix of size from left to right $10, 50, 200$ and $\gamma = 0.6$ and the Tracy-Widom law.

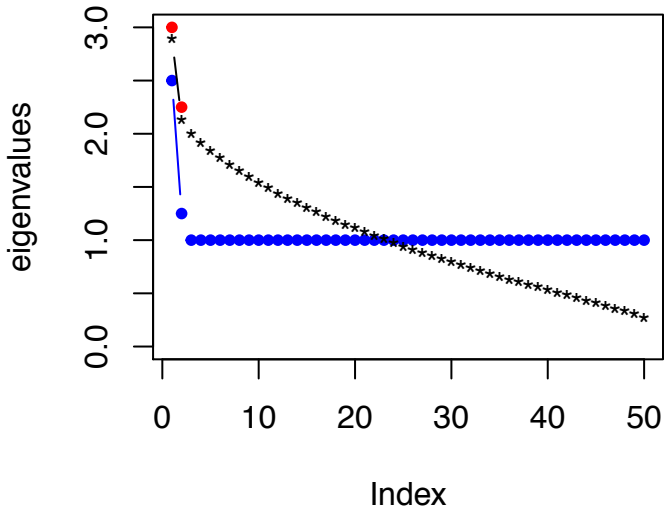

Figure 6: Wishart matrix with $\gamma = 0.25$ under the spiked model. True eigenvalues (blue), empirical eigenvalues (black) and limiting two first eigenvalues (red) of the spiked model with $l_1 = 2 > 1+\sqrt{\gamma}$ and $l_2 = 1.25 \leq 1 + \sqrt{\gamma}$.

$\hat{l}_1, \ldots, \hat{l}_M$ the M largest eigenvalues of S the empirical covariance matrix of X. Assume $p, n \to \infty$ such that $p/n \to \gamma \in (0, \infty)$. Then for all $j = 1, \ldots, M$

$$\hat{l}_j \overset{a.s.}{\to} \begin{cases} (1+\sqrt{\gamma})^2 & \text{if} \quad l_j \leq 1+\sqrt{\gamma} \\ l_j(1+\gamma/(l_j-1)) & \text{otherwise.} \end{cases}$$

Figure 6 illustrates this result with a Gaussian matrix under the spiked model. The covariance matrix is defined as $\Sigma = l_1\theta_1\theta_1^T + l_2\theta_2\theta_2^T + I_p$, $n = 200$ and $p = 50$. Thus $\gamma = 0.25$. The first eigenvalue $l_1 = 2$ is over the threshold $1 + \sqrt{\gamma} = 1.5$ and the second eigenvalue $l_2 = 1.25$ is under the threshold $1 + \sqrt{\gamma} = 1.5$. The plot represents the eigenvalues of the sample covariance matrix S (black), the true eigenvalues of Σ (blue) and the limit of the two first eigenvalues of S given by the theorem.

4.3 MANOVA

Consider g groups of multivariate normal vectors $X^{(1)}, \ldots, X^{(g)}$ with distributions $\mathcal{N}(\mu_i, \Sigma)$. The $X^{(i)}$ have a common covariance matrix Σ and we want to test the null hypothesis $H_0 : \mu_1 = \ldots = \mu_g$ agaisnt the alternative $H_1 : \mu_i \neq \mu_j$ for some i, j. Assume that we observe samples of the $X^{(i)}$ of sizes n_i respectively. $\bar{X}^{(i)}$ is the mean of the group $X^{(i)}$ while $\bar{X}$ is the overall mean. Define the between sample covariance matrix $U = \sum_{i=1}^{g} n_i(\bar{X}^{(i)} - \bar{X})(\bar{X}^{(i)} - \bar{X})^T$ and the within sample covariance matrix $V = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_j^{(i)} - \bar{X}^{(i)})(X_j^{(i)} - \bar{X}^{(i)})^T$. Under H_0, U and V are independent Wishart matrices and thus the F-matrix UV^{-1} is a double Wishart. Define l_1 the largest eigenvalue

of UV^{-1}. Asymptotically $(\log(l_1/1-l_1)-\mu_p)/\sigma_p$ follows the TW_1 distribution. An asymptotic test of level $1-\varepsilon$ is to reject H_0 if $(\log(l_1/1-l_1)-\mu_p)/\sigma_p$ is above the $(1-\varepsilon)$-quantile of TW_1 and to accept it otherwise.

5 Conclusion

In this document I gave the most classical results reviewed by Aue & Paul. Many other results can be found in their article. Some consider for example non i.i.d. settings, others concern the behavior of eigenvectors.