

# The Distributions of Random Matrix Theory and their Applications\*

Craig A. Tracy<sup>†</sup> and Harold Widom<sup>‡</sup>

## Abstract

This paper surveys the largest eigenvalue distributions appearing in random matrix theory and their application to multivariate statistical analysis.

## Contents

<b>1</b>	<b>Random Matrix Models: Gaussian Ensembles</b>	<b>2</b>
1.1	Largest eigenvalue distributions $F_\beta$ . Painlevé II Representations . . . . .	2
1.1.1	Tail behavior of $F_\beta$ . . . . .	3
1.1.2	Numerical evaluation of $F_\beta$ . . . . .	5
1.2	Next-largest, next-next largest, etc. eigenvalue distributions . . . . .	5
<b>2</b>	<b>Universality Theorems</b>	<b>5</b>
2.1	Invariant Ensembles . . . . .	5
2.2	Wigner Ensembles . . . . .	6
<b>3</b>	<b>Multivariate Statistical Analysis</b>	<b>7</b>
3.1	Principal Component Analysis (PCA) . . . . .	7
3.2	Testing the Null Hypothesis . . . . .	8
3.3	Spiked Populations: BBP Phase Transition . . . . .	9
<b>4</b>	<b>Conclusions</b>	<b>11</b>

---

\*Submitted to the Stanford Institute for Theoretical Economics Summer 2008 Workshop: *Complex Data in Economics and Finance: Spatial Models, Social Networks and Factor Models*. This work was supported by the National Science Foundation under grants DMS-0553379 (first author) and DMS-0552388 (second author).

<sup>†</sup>Department of Mathematics, University of California, Davis, CA 95616. Email: tracy@math.ucdavis.edu.

<sup>‡</sup>Department of Mathematics, University of California, Santa Cruz, CA 95064. Email: widom@ucsc.edu.

# 1 Random Matrix Models: Gaussian Ensembles

A random matrix model (RMM) is a probability space  $(\Omega, \mathbb{P}, \mathcal{F})$  where the sample space  $\Omega$  is a set of matrices. There are three classic finite- $N$  RMM called the *Gaussian ensembles* (see, e.g. [23] and for early history [30]):

- Gaussian Orthogonal Ensemble (GOE,  $\beta = 1$ )
  - $\Omega = N \times N$  real symmetric matrices
  - $\mathbb{P} =$  unique (up to a choice of the mean and variance) measure that is invariant under orthogonal transformations and the algebraically independent matrix elements are i.i.d. random variables. Explicitly (for mean zero and a choice of the variance), the density is

$$c_N \exp(-\text{tr}(A^2)/2) dA, \tag{1}$$

where  $c_N$  is a normalization constant and  $dA = \prod_i dA_{ii} \prod_{i < j} dA_{ij}$ , the product Lebesgue measure on the algebraically independent matrix elements.

- Gaussian Unitary Ensemble (GUE,  $\beta = 2$ )
  - $\Omega = N \times N$  hermitian matrices
  - $\mathbb{P} =$  unique measure (again up to a choice of the mean and variance) that is invariant under unitary transformations and the algebraically independent real and imaginary matrix elements are i.i.d. random variables. Again the density is of the form (1) with  $dA = \prod_i dA_{ii} \prod_{i < j} d\Re(A_{ij}) d\Im(A_{ij})$ .
- Gaussian Symplectic Ensemble (GSE,  $\beta = 4$ ) (see [23] for a definition)

For  $A$  in any of the above Gaussian ensembles, let  $\lambda_1(A) \leq \dots \leq \lambda_N(A) := \lambda_{\max}$  denote the eigenvalues of  $A$ . These eigenvalues are real and define random variables on the respective probability spaces. (With probability one the eigenvalues are distinct.) Since these Gaussian ensembles are defined by invariant measures, one can explicitly compute the joint distribution of eigenvalues and show that it has the following density with respect to Lebesgue measure:

$$\mathbb{P}_{\beta, N}(x_1, \dots, x_N) = C_{N, \beta} \prod_{1 \leq i < j \leq N} |x_i - x_j|^\beta \prod_{i=1}^N e^{-\beta x_i^2/2}, \quad \beta = 1, 2, 4,$$

where  $C_{N, \beta}$  is a known normalization constant [23]. The form of the joint density explains the usefulness of the  $\beta$  notation.

## 1.1 Largest eigenvalue distributions $F_\beta$ . Painlevé II Representations

Generally speaking, the interest lies in limit laws as  $N \rightarrow \infty$ . As is familiar from the central limit theorem, to get nontrivial limits one must center and normalize the random variables. Here the main focus is on the limit law associated with the largest eigenvalue. If

$$F_{N, \beta}(t) := \mathbb{P}_{\beta, N}(\lambda_{\max} < t), \quad \beta = 1, 2, 4,$$

denotes the distribution function of the largest eigenvalue, then the basic limit laws [36, 37, 38] state that<sup>1</sup>

$$F_\beta(x) := \lim_{N \rightarrow \infty} F_{N,\beta} \left( 2\sigma\sqrt{N} + \frac{\sigma x}{N^{1/6}} \right), \quad \beta = 1, 2, 4,$$

exist and are given explicitly by

$$F_2(x) = \exp \left( - \int_x^\infty (y-x)q^2(y) dy \right) \quad (2)$$

where  $q$  is the unique solution<sup>2</sup> to the *Painlevé II equation*

$$\frac{d^2q}{dx^2} = xq + 2q^3$$

satisfying the boundary condition<sup>3</sup>

$$q(x) \sim \text{Ai}(x) \text{ as } x \rightarrow \infty. \quad (3)$$

It is known [17] that

$$q(x) = \sqrt{-\frac{x}{2}} \left( 1 + \frac{1}{8x^3} + O\left(\frac{1}{x^6}\right) \right) \text{ as } x \rightarrow -\infty.$$

The orthogonal and symplectic distributions [38] are

$$F_1(x) = \exp \left( -\frac{1}{2} \int_x^\infty q(y) dy \right) (F_2(x))^{1/2}, \quad (4)$$

$$F_4(x/\sqrt{2}) = \cosh \left( \frac{1}{2} \int_x^\infty q(y) dy \right) (F_2(x))^{1/2}. \quad (5)$$

Graphs of the densities  $f_\beta := dF_\beta/dx$  are in Figure 1 and some statistics of  $F_\beta$  can be found in the Table 1.

### 1.1.1 Tail behavior of $F_\beta$

The asymptotics for  $F_\beta(x)$  as  $x \rightarrow +\infty$  follows straightforwardly from (2)–(5). To state the results it is first convenient to introduce

$$\begin{aligned} F(x) &= \exp \left( -\frac{1}{2} \int_x^\infty (y-x)q(y)^2 dy \right), \\ E(x) &= \exp \left( -\frac{1}{2} \int_x^\infty q(y) dy \right) \end{aligned}$$

---

<sup>1</sup>Here  $\sigma$  is the standard deviation of the Gaussian distribution on the off-diagonal matrix elements. For the normalization we've chosen,  $\sigma = 1/\sqrt{2}$ ; however, other choices are common.

<sup>2</sup>That such a unique solution exists is a nontrivial fact first proved by Hastings and McLeod [17]; and for this reason,  $q$  is often called the Hastings-McLeod solution. See [13] for a detailed account of Painlevé transcendents.

<sup>3</sup>Ai is the Airy function.

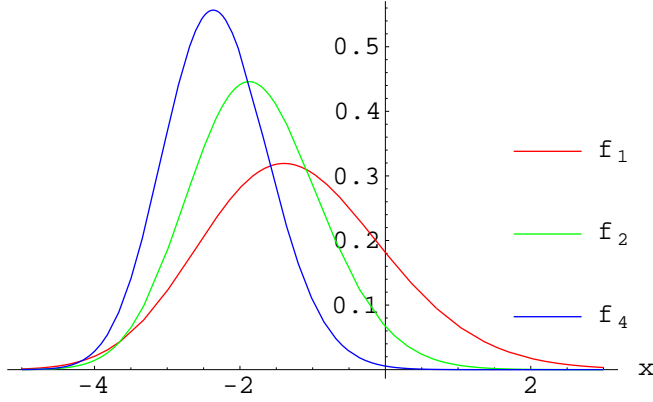


Figure 1: Largest eigenvalue densities  $f_\beta(x) = dF_\beta/dx$ ,  $\beta = 1, 2, 4$  where  $F_\beta$  are defined in (2), (4) and (5).

so that

$$F_1(x) = E(x)F(x), \quad F_2(x) = F(x)^2, \quad \text{and} \quad F_4(x/\sqrt{2}) = \frac{1}{2} \left( E(x) + \frac{1}{E(x)} \right) F(x).$$

Then as  $x \rightarrow +\infty$

$$F(x) = 1 - \frac{e^{-\frac{4}{3}x^{3/2}}}{32\pi x^{3/2}} \left( 1 + O\left(\frac{1}{x^{3/2}}\right) \right),$$

$$E(x) = 1 - \frac{e^{-\frac{2}{3}x^{3/2}}}{4\sqrt{\pi}x^{3/2}} \left( 1 + O\left(\frac{1}{x^{3/2}}\right) \right)$$

from which the asymptotics for  $F_\beta$  follows.

The asymptotics as  $x \rightarrow -\infty$  is much more difficult and the complete solution was only recently achieved for  $\beta = 1, 2, 4$  [3]. We quote the final results and refer the reader to [3] for a history of this problem. As  $x \rightarrow -\infty$

$$F_1(x) = \tau_1 \frac{e^{-\frac{1}{24}|x|^3 - \frac{1}{3\sqrt{2}}|x|^{3/2}}}{|x|^{1/16}} \left( 1 - \frac{1}{24\sqrt{2}|x|^{3/2}} + O(|x|^{-3}) \right),$$

$$F_2(x) = \tau_2 \frac{e^{-\frac{1}{12}|x|^3}}{|x|^{1/8}} \left( 1 + \frac{3}{26|x|^3} + O(|x|^{-6}) \right),$$

$$F_4(x/\sqrt{2}) = \tau_4 \frac{e^{-\frac{1}{24}|x|^3 + \frac{1}{3\sqrt{2}}|x|^{3/2}}}{|x|^{1/16}} \left( 1 + \frac{1}{24\sqrt{2}|x|^{3/2}} + O(|x|^{-3}) \right)$$

where

$$\tau_1 = 2^{-11/48} e^{\frac{1}{2}\zeta'(-1)}, \quad \tau_2 = 2^{1/24} e^{\zeta'(-1)}, \quad \tau_4 = 2^{-35/48} e^{\frac{1}{2}\zeta'(-1)}$$

and  $\zeta'(-1) = -0.1654211437\dots$  is the derivative of the Riemann zeta function evaluated at  $-1$ .

Table 1: The mean ( $\mu_\beta$ ), variance ( $\sigma_\beta^2$ ), skewness ( $S_\beta$ ) and kurtosis ( $K_\beta$ ) of  $F_\beta$ . The high-precision numbers are courtesy of Michael Prähofer.

$\beta$	$\mu_\beta$	$\sigma_\beta^2$	$S_\beta$	$K_\beta$
1	-1.206 533 574	1.607 781 034	0.293 464 524	0.165 242 938
2	-1.771 086 807	0.813 194 792	0.224 084 203	0.093 448 087
4	-2.3069	0.5177	0.1655	0.0492

### 1.1.2 Numerical evaluation of $F_\beta$

Particularly for applications to data analysis, it is useful to have numerical evaluations of the distributions  $F_\beta$ . Chapter 7 of Dieng’s Ph.D. thesis [12] gives MATLAB<sup>TM</sup> code to evaluate and plot these distributions. Tables of the Hastings-McLeod solution to Painlevé II and  $F_{1,2}$  can be found on Prähofer’s homepage [31].<sup>4</sup> A different approach [6] to the numerical evaluation of  $F_\beta$  is based on the Fredholm determinant representations for  $F_\beta$  (see, e.g. [39]).

## 1.2 Next-largest, next-next largest, etc. eigenvalue distributions

There exist Painlevé II type representations for the limiting distributions of the next-largest eigenvalue ( $\lambda_{N-1}$ ), next-next largest eigenvalue ( $\lambda_{N-2}$ ), etc. The unitary case was examined some time ago [37] but only recently did Dieng [11] derive limiting distributions for the orthogonal and symplectic cases. It should be remarked that the results in the orthogonal case were somewhat surprising. Figure 2 displays simulations for the four largest eigenvalues of  $N = 1000$  GOE matrices and their respective limiting distributions.

## 2 Universality Theorems

A natural question is to what extent do the above limit laws depend upon the Gaussian and invariance assumptions for the probability measure?

### 2.1 Invariant Ensembles

A more general class of invariant RMM results by replacing the Gaussian measures with

$$d\mathbb{P}_N(A) = c_{N,\beta} \exp(-\beta \text{tr}(V(A))/2) dA$$

where  $V$  is a polynomial of even degree and positive leading coefficient. This implies that the joint density for the eigenvalues is

$$\mathbb{P}_{\beta,V,N}(x_1, \dots, x_N) = C_{V,N,\beta} \prod_{1 \leq i < j \leq N} |x_i - x_j|^\beta \prod_{i=1}^N e^{-\beta V(x_i)/2}, \quad \beta = 1, 2, 4, \quad (6)$$

---

<sup>4</sup>Note that the Hastings-McLeod solution in the Prähofer tables is denoted  $u(s)$  and in the notation here  $u(s) = -q(s)$ .

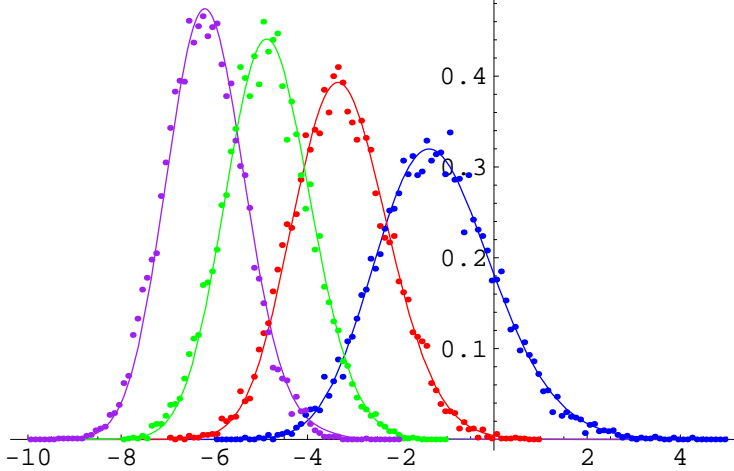


Figure 2: A histogram of the four largest (centered and normalized) eigenvalues for  $10^4$  realizations of  $10^3 \times 10^3$  GOE matrices. Solid curves are the limiting distributions from [11]. Figure a courtesy of Momar Dieng.

where  $C_{V,N,\beta}$  is a normalization constant [23]. Unitary ensembles ( $\beta = 2$ ) are technically simpler than the orthogonal and symplectic ensembles ( $\beta = 1, 4$ ), but both require for general  $V$  powerful Riemann-Hilbert methods [10] for the asymptotic analysis. The main conclusions from these studies for the limiting distribution of the largest eigenvalue are

**Theorem.** There exist constants  $z_N^{(\beta)}$  and  $s_N^{(\beta)}$  such that

$$\lim_{N \rightarrow \infty} \mathbb{P}_{\beta,V,N} \left( \frac{\lambda_{\max} - z_N^{(\beta)}}{s_N^{(\beta)}} \leq t \right) = F_{\beta}(t), \quad \beta = 1, 2, 4,$$

where the  $F_{\beta}$  are given by (2), (4) and (5).

The results for the unitary case ( $\beta = 2$ ) are due to Deift, Kriecherbaur, McLaughlin, Venakides and Zhou [9] and the orthogonal/symplectic results are recent work of Deift and Gioev [8]. The universality theorem for special case  $V(A) = \frac{1}{4}A^4 - gA^2$  is due to Bleher and Its [5] ( $\beta = 2$ ) and Stojanovic [35] ( $\beta = 1$ ). These deep theorems broadly extend the domain of attraction of the  $F_{\beta}$  limit laws. Deift's ICM 2006 lecture [7] is a recommended overview for these developments.

## 2.2 Wigner Ensembles

Wigner matrices are RMM of complex hermitian or real symmetric  $N \times N$  matrices  $H$

$$H = \frac{1}{\sqrt{N}}(A_{ij})_{i,j=1}^N$$

where  $A_{ij}$ ,  $1 \leq i < j \leq N$  are i.i.d. complex or real random variables with distribution  $\mu$ . The diagonal matrix elements are i.i.d. real random variables independent of the off-diagonal elements. The diagonal probability distribution is centered, independent of  $N$  and has finite variance. They are called *Wigner matrices* since Wigner in 1955 first studied the limiting distribution of the empirical spectral measure under the assumption that  $\mu$  has

finite variance. The limiting spectral measure is the famous *Wigner semicircle* distribution. We denote the Wigner measure on the space of either complex Hermitian or real symmetric  $N \times N$  matrices by  $\mathbb{P}_{W,N}$

Except in the case of the Gaussian distribution, the Wigner ensembles define non-invariant measures. For this reason no explicit formulas for the joint distribution of eigenvalues, such as (6) for invariant measures, are known. Thus the techniques used to prove universality theorems have a completely different flavor.

Soshnikov [32] proved, under the additional assumptions that  $\mu$  is symmetric (all odd moments are zero) and the distribution decays as at least as fast as a Gaussian distribution together with a normalization on the variances,<sup>5</sup> the following universality statement for the largest eigenvalue  $\lambda_{\max}$  of Wigner random matrices

**Theorem.**

$$\lim_{N \rightarrow \infty} \mathbb{P}_{W,N} \left( \lambda_{\max} \leq 1 + \frac{x}{2N^{2/3}} \right) = F_{\beta}(x)$$

with  $\beta = 1$  for real symmetric matrices and  $\beta = 2$  for complex hermitian matrices.

The importance of Soshnikov’s theorem is the universality of  $F_{\beta}$  has been established for ensembles for which the “integrable” techniques, e.g. Fredholm theory, Riemann-Hilbert methods, Painlevé theory, are not directly applicable. Current research [29] is exploring the relaxation of the symmetry constraint on the underlying distribution  $\mu$ .

### 3 Multivariate Statistical Analysis

As Johnstone [22] remarked:

It is a striking feature of the classical theory of multivariate statistical analysis that most of the standard techniques—principal components, canonical correlations, multivariate analysis of variance (MANOVA), discriminant analysis and so forth—are founded on the eigenanalysis of covariance matrices.

Thus it is not surprising that the methods of random matrix theory have important applications to multivariate statistical analysis. We now survey some of these recent developments drawing heavily on Johnstone’s 2006 ICM lecture [21]. We have also benefited from the unpublished survey by Pécché [28].

#### 3.1 Principal Component Analysis (PCA)

Recall that in PCA with  $p$  variables one distinguishes between the *population eigenvalues*  $\ell_j$ , which are the eigenvalues of the underlying  $p \times p$  covariance matrix

$$\Sigma = (\text{Cov}(X_k, X_{k'}))_{1 \leq k, k' \leq p},$$

and the *sample eigenvalues*  $\hat{\ell}_j$ , which are the (random) eigenvalues of the sample covariance matrix

$$S = \frac{1}{n} X X^T.$$

---

<sup>5</sup>For real symmetric matrices the normalization is  $E_{W,N}(H_{ij}^2) = \frac{1}{4}$ ,  $1 \leq i < j \leq N$  and for complex hermitian matrices  $\mathbb{E}_{W,N}(\Re(H_{ij})^2) = \mathbb{E}_{W,N}(\Im(H_{ij})^2) = \frac{1}{8}$ .

Here  $X$  is the  $p \times n$  data matrix and  $n$  is the number of observations of the  $p$  variables. (A column of  $X$  represents one observation of the  $p$  variables.) Since the parameters of the underlying probability model describing the random variables  $X_1, \dots, X_p$  are unknown, the problem is to deduce properties of  $\Sigma$  from the observed sample covariance matrix  $S$ .

The simplest model is to assume  $\mathbb{X} = (X_1, \dots, X_p)$  is a  $p$ -variate Gaussian distribution  $N_p(\mu, \Sigma)$  and the data matrix  $X$  is formed by  $n$  independent draws  $\mathbb{X}_1, \dots, \mathbb{X}_n$ . (For simplicity we consider  $\mu = 0$ .) The  $p \times p$  matrix  $A = XX^T$  is said to have  $p$ -variate *Wishart distribution* on  $n$  degrees of freedom,  $W_p(n, \Sigma)$ . We denote the eigenvalues of  $A$  by  $l_1 \geq l_2 \geq \dots \geq l_p \geq 0$  (so  $l_j = n\hat{\ell}_j$ ). The joint distribution of the eigenvalues  $l_j$  has been known for some time (e.g. Muirhead [24], Theorem 3.2.18) and is complicated by the fact it involves an integral over the orthogonal group  $\mathbb{O}(p)$ .

### 3.2 Testing the Null Hypothesis

The null hypothesis  $H_0$  is the statement that there are no correlations amongst the  $p$  variables, i.e.  $\Sigma = I$ . Under  $H_0$  all the population eigenvalues equal one, but as been known for some time<sup>6</sup> there is a “spread” in the sample eigenvalues  $\hat{\ell}_j$ . To assess whether “large” observed eigenvalues justify rejecting the null hypothesis, we need an approximation to the the *null hypothesis distribution* of the largest sample eigenvalue,

$$\mathbb{P}\left(\hat{\ell}_1 > t | H_0 = W_p(n, I)\right). \quad (7)$$

This approximation is provided by the following theorem of Johnstone [20].

**Theorem.**

$$\mathbb{P}\left(n\hat{\ell}_1 \leq \mu_{np} + \sigma_{np}x | H_0\right) \longrightarrow F_1(x)$$

where the limit is  $n \rightarrow \infty$ ,  $p \rightarrow \infty$  such that  $p/n \rightarrow \gamma \in (0, \infty)$ ,  $F_1$  is the largest eigenvalue distribution (4), and the centering and norming constants are

$$\mu_{np} = \left(\sqrt{n - \frac{1}{2}} + \sqrt{p - \frac{1}{2}}\right)^2, \quad (8)$$

$$\sigma_{np} = (\sqrt{n} + \sqrt{p}) \left(\frac{1}{\sqrt{n - \frac{1}{2}}} + \frac{1}{\sqrt{p - \frac{1}{2}}}\right)^{1/3}. \quad (9)$$

Several remarks are in order.

1. The appearance of the fractions  $\frac{1}{2}$  in  $\mu_{np}$  and  $\sigma_{np}$  appear to improve the rate of convergence to  $F_1$  to “second-order accuracy” [21]. With this choice of constants,  $F_1$  provides a good approximation for rather small values of  $p$ . (See Johnstone’s comparisons with the tables of Chen [21].)
2. El Karoui [14] shows the theorem holds more generally as

$$p/n \rightarrow \gamma \in [0, \infty].$$

---

<sup>6</sup>For  $\Sigma = I$ , the density of eigenvalues of  $S$  follows the Marčenko-Pastur distribution, a generalization of the Wigner semicircle distribution.

Table 2: Values of  $x$  for given  $\mathbb{P}(\chi_1 \geq x)$  where  $\chi_1$  has distribution  $F_1$ .

$x$	$\mathbb{P}(\chi_1 \geq x)$
2.02345	.01
1.59776	.02
1.33321	.03
1.13706	.04
0.97931	.05
0.84633	.06
0.73069	.07
0.62792	.08
0.53508	.09
0.45014	.10

3. For complex data matrices with  $\Sigma = I$ , there are corresponding limit theorems where now convergence is to  $F_2$  [18, 20].
4. Soshnikov [33] and P  ch   [27] have removed the assumption of Gaussian samples. They assume that the matrix elements  $X_{ij}$  of the data matrix  $X$  are independent random variables with a common symmetric distribution whose moments grow not faster than the Gaussian ones. We refer the reader to [27] for a description of the centering and norming constants. Limit theorems for complex data matrices are also proved.
5. To summarize, given the centering and norming constants (8) and (9) together with tables such as Table 2, one has a good approximation to the null distribution function (7).

### 3.3 Spiked Populations: BBP Phase Transition

As mentioned above, an essential difficulty in extending the above limit laws for  $\hat{\ell}_1$  when the  $A = XX^T \in W_p(n, \Sigma)$ ,  $\Sigma \neq I$ , is the presence of a certain integral over the orthogonal group  $\mathbb{O}(p)$  in the joint distribution of eigenvalues of  $A$ . In the case of *complex* Wishart matrices, the corresponding integral in the joint distribution of eigenvalues is over the unitary group  $\mathbb{U}(p)$  which, fortunately, can be explicitly evaluated by use of the Harish-Chandra-Itzykson-Zuber formula, see, e.g. [40].

We now describe the limit theorem of Baik, Ben Arous and P  ch   [2] where they consider the *complex* Wishart ensemble with the  $p \times p$  covariance matrix

$$\Sigma = \text{diag}(\ell_1, \dots, \ell_r, 1, \dots, 1).$$

For ease of exposition of their results, we consider  $r = 1$  with  $\ell_1 > 1$ . As before we consider the limit

$$p \rightarrow \infty, \quad n \rightarrow \infty \quad \text{such that} \quad \frac{p}{n} \rightarrow \gamma \geq 1. \quad (10)$$

Define

$$w_c = 1 + \sqrt{\gamma}.$$

**Theorem.** With  $\Sigma$  as above ( $r = 1$ ), let  $\hat{\ell}_1$  denote the largest eigenvalue of the sample covariance matrix.

- If  $1 \leq \ell_1 < w_c$ , then in the limit (10)

$$\mathbb{P}\left(\frac{n^{2/3}}{\sigma}(\hat{\ell}_1 - \mu) \leq x\right) \rightarrow F_2(x),$$

where  $F_2$  is given by (2) and

$$\mu = (1 + \sqrt{\gamma})^2, \quad \sigma = (1 + \sqrt{\gamma})\left(1 + \frac{1}{\sqrt{\gamma}}\right)^{1/3}.$$

- If  $\pi_1 > w_c$ , then in the limit (10)

$$\mathbb{P}\left(\frac{n^{1/2}}{\sigma_1}(\hat{\ell}_1 - \mu_1) \leq x\right) \rightarrow \Phi(x),$$

where  $\Phi$  is the standard normal distribution and

$$\mu_1 = \ell_1 \left(1 + \frac{\gamma}{\ell_1 - 1}\right), \quad \sigma_1 = \ell_1^2 \left(1 - \frac{\gamma}{(\ell_1 - 1)^2}\right).$$

Remarks:

1. The BBP theorem “shows that a single eigenvalue of the true covariance  $\Sigma$  may drastically change the limiting behavior of the largest eigenvalue of sample covariance matrices. One should understand the above result as the statement that the eigenvalues exiting the support of the Marchenko-Pastur distribution form a small bulk of eigenvalues. This small bulk exhibits the same eigenvalue statistics as the eigenvalues of a non-normalized GUE (resp. GOE) matrix” [28].
2. If  $\pi_1 = w_c$  the limiting distribution is a generalization of  $F_2$  expressible in terms of the same Painlevé II function  $q$  [1].
3. For real Wishart matrices, Paul [25] shows that if  $\pi_1 > w_c$  is simple, then  $\hat{\ell}_1$  exhibits Gaussian fluctuations.
4. El Karoui [15] finds a large class of complex Wishart matrices  $W_p(\Sigma, n)$  which have a  $F_2$  limit law for  $\hat{\ell}_1$ .
5. Patterson, Price and Reich [26] have applied these results to problems of population structure arising from genetic data. See Harding [16] for an application in economics.

## 4 Conclusions

In this note we have surveyed some basic properties of the largest eigenvalue distributions  $F_\beta$ , their appearance as limit laws for large classes of random matrix models as well as their application to principal component analysis. We mention that these same distributions play an analogous role in canonical correlations [22] as they do in PCA. Though not discussed in these notes, the same  $F_\beta$  appear as limit laws for certain problems in combinatorial theory related to growth processes [4, 18]. (For a recent review of these topics see [34].)

## References

- [1] J. Baik, Painlevé formulas of the limiting distributions for nonnull complex sample covariance matrices of spiked population models, *Duke Math. J.* **133** (2006), 205–235.
- [2] J. Baik, G. Ben Arous and S. Péché, Phase transition of the largest eigenvalue for non-null complex sample covariance matrices, *Ann. of Probab.* **33** (2005), 1643–1697.
- [3] J. Baik, R. Buckingham and J. DiFranco, Asymptotics of the Tracy-Widom distributions and the total integral of a Painlevé II function, *Commun. Math. Phys.* **280** (2008), 463–497.
- [4] J. Baik, P. Deift and K. Johansson, On the distribution of the length of the longest increasing subsequence of random permutations, *J. Amer. Math. Soc.* **12** (1999), 1119–1178.
- [5] P. Bleher and A. Its, Semiclassical asymptotics of orthogonal polynomials, Riemann-Hilbert problem, and universality in the matrix model, *Ann. Math.* **150** (1999), 185–266.
- [6] F. Bornemann, On the numerical evaluation of Fredholm determinants, arXiv:0804.2543.
- [7] P. Deift, Universality for mathematical and physical systems, *International Congress of Mathematicians*, Vol. I, 125–152, Eur. Math. Soc., Zürich, 2007. Available in preprint form at <http://front.math.ucdavis.edu/0603.4738>
- [8] P. Deift and D. Gioev, Universality at the edge of the spectrum for unitary, orthogonal, and symplectic ensembles of random matrices, *Comm. Pure Appl. Math.* **60** (2007), 867–910.
- [9] P. Deift, T. Kriecherbauer, K. T-R. McLaughlin, S. Venakides and X. Zhou, Uniform asymptotics for polynomials orthogonal with respect to varying exponential weight and applications to universality questions in random matrix theory, *Comm. Pure Appl. Math.* **52** (1999), 1335–1425.
- [10] P. A. Deift and X. Zhou, A steepest descent method for oscillatory Riemann-Hilbert problems. Asymptotics for the MKdV equation, *Ann. of Math.* **137** (1993), 295–368.
- [11] M. Dieng, Distribution functions for edge eigenvalues in orthogonal and symplectic ensembles: Painlevé representations, *Int. Math. Res. Not.* **37** (2005), 2263–2287.

- [12] M. Dieng, Distribution functions for edge eigenvalues in orthogonal and symplectic ensembles: Painlevé Representations II, arXiv:math/0506586.
- [13] A. S. Fokas, A. R. Its, A. A. Kapaev and V. Yu. Novokshenov, *Painlevé Transcendents: The Riemann-Hilbert Approach*, American Mathematical Society, 2006.
- [14] N. El Karoui, On the largest eigenvalue of Wishart matrices with identity covariance when  $n$ ,  $p$  and  $p/n$  tend to infinity, arXiv:math/0309355.
- [15] N. El Karoui, Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices, *Ann. of Probab.* **35** (2007), 663–714.
- [16] M. Harding, Explaining the single factor bias of arbitrage pricing models in finite samples, *Economics Letters* **99** (2008), 85–88.
- [17] S. P. Hastings and J. B. McLeod, A boundary value problem associated with the second Painlevé transcendent and the Korteweg-de Vries equation, *Arch. Rational Mech. Anal.* **73** (1980), 31–51.
- [18] K. Johansson, Shape fluctuations and random matrices, *Commun. Math. Phys.* **209** (2000), 437–476.
- [19] K. Johansson, Toeplitz determinants, random growth and determinantal processes, ICM 2002, Vol. III, 53–62.
- [20] I. Johnstone, On the distribution of the largest principal component, *Ann. Statistics* **29** (2001), 295–327.
- [21] I. Johnstone, High dimensional statistical inference and random matrices, *International Congress of Mathematicians*, Vol. I, 307–333, Eur. Math. Soc. Zürich, 2007. Available in preprint form at <http://front.math.ucdavis.edu/0611.5589>
- [22] I. Johnstone, Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy Widom limits and rates of convergence, preprint, arXiv: 0803.3408.
- [23] M. L. Mehta, *Random Matrices*, second ed., Academic Press, 1991.
- [24] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, 1982.
- [25] D. Paul, Asymptotics of the leading sample eigenvalues for a spiked covariance model, *Stat. Sinica* **17** (2007), 1617–1642.
- [26] N. Patterson, A. L. Price and D. Reich, Population structure and eigenanalysis, *PLoS Genetics* **2(12)** (2006): e190.
- [27] S. Péché, Universality results for largest eigenvalues of some sample covariance matrices, arXiv:0705.1701.
- [28] S. Péché, The edge of the spectrum of random matrices, *Habilitation à diriger des recherches*, Université Joseph Fourier Grenoble I, to be submitted.

- [29] S. Péché and A. Soshnikov, On the lower bound of the spectral norm of symmetric random matrices with independent entries, *Elect. Comm. in Probab.* **13** (2008), 280–290.
- [30] C. E. Porter, *Statistical Theories of Spectra: Fluctuations*, Academic Press, 1965.
- [31] M. Prähofer, <http://www-m5.ma.tum.de/pers/praehofer/>
- [32] A. Soshnikov, Universality at the edge of the spectrum in Wigner random matrices, *Commun. Math. Phys.* **207** (1999), 697–733.
- [33] A. Soshnikov, A note on universality of the distribution of the largest eigenvalue in certain classes of sample covariance matrices, *J. Stat. Phys.* **108** (2002), 1033–1056.
- [34] H. Spohn, Exact solutions for KPZ-type growth processes, random matrices, and equilibrium shapes of crystals, *Physica A* **369** (2006), 71–99.
- [35] A. Stojanovic, Universality in orthogonal and symplectic invariant matrix models with quartic potential, *Math. Phys., Anal. and Geom.* **3** (2000), 339–373.
- [36] C. A. Tracy and H. Widom, Level-spacing distribution and the Airy kernel, *Phys. Letts. B* **305** (1993), 115–118.
- [37] C. A. Tracy and H. Widom, Level-spacing distribution and the Airy kernel, *Commun. Math. Phys.* **159** (1994), 151–174.
- [38] C. A. Tracy and H. Widom, On orthogonal and symplectic matrix ensembles, *Commun. Math. Phys.* **177** (1996), 727–754.
- [39] C. A. Tracy and H. Widom, Matrix kernels for the Gaussian orthogonal and symplectic ensembles, *Ann. Inst. Fourier, Grenoble* **55** (2005), 2197–2207.
- [40] P. Zinn-Justin and J.-B. Zuber, On some integrals over the  $U(N)$  unitary group and their large  $N$  limit, *J. Phys. A: Math. Gen.* **36** (2003), 3173–3193.