An introduction to random matrix theory

Gaëtan Borot[*]

Mini-course, National University of Mongolia, Ulaanbaatar

August 7th, 2015

Contents

[*]Max-Planck Institut for Mathematics: gborot@mpim-bonn.mpg.de

These are lectures notes for a 4h30 mini-course held in Ulaanbaatar, National University of Mongolia, August 5-7th 2015, at the summer school

Stochastic Processes and Applications, Mongolia

I thank Carina Geldhauser, Andreas Kyprianou, Tsogzolmaa Saizmaa and the local organizers in Mongolia to have arranged this event, as well as the DAAD, the University of Augsburg and Lisa Beck for funding.

The aim is to present an introduction to basic results of random matrix theory and some of its motivations, targeted to a large panel of students coming from statistics, finance, etc. Only a small background in probability is required (Mongolian students had a 1.5 month crash course on measure theory before the summer school). A few references to support – or go further than – the course:

- *High Dimensional Statistical Inference and Random Matrices*, I. Johnstone, Proceedings of the ICM, Madrid, Spain, (2006), math.ST/0611589. A short review of the application of random matrix theory results to statistics.
- *Theory of finance risks: from statistical physics to risk management*, J.P. Bouchaud and M. Potters, CUP (2000). A book explaining how ideas coming from statistical physics (and for a small part, of random matrices) can be applied to finance, by two pioneers. J.P. Bouchaud founded a hedge fund (Capital Fund Management), which conduct investment using those ideas, as well as pure research.
- *Population structure and eigenanalysis*, N. Patterson, A.L. Preis and D. Reich, PLoS Genetics **2** 12 (2006). Research discussing the methodology of PCA, and proposing statistical tests based on Tracy-Widom distributions, with applications to population genetics in view.
- *Random matrices*, M.L. Mehta, 3rd edition, Elsevier (2004). Written by a pioneer of random matrix theory. Accessible at master level, rather focused on calculations and results for exactly solvable models, including Gaussian ensembles. A good reference to browse for results.

Lecture 1 (1h30)
August 5th, 2015

1 Motivations from statistics for data in high dimensions

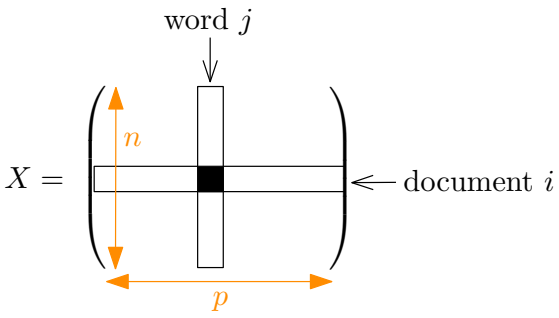
Collecting a huge amount of data has been facilitated by the development of computer sciences. It is then critical to have tools to analyze this data. Imagine that for each sample one has collected information represented by a point in $\mathbb{R}^N$. With N large, this is certainly too much information for our brain to process. One would like to know if some relevant patterns can be identified, that would explain most of the scattering of the data by restricting to a well-chosen k-dimensional plane in $\mathbb{R}^N$, for $k = 1, 2, 3$ etc. This problem is posed for instance in archeology, in biology and genetics, in economics and finance, in linguistics, etc. Let us give some examples.

1.1 *Latent semantics*

Imagine we have documents $i \in \{1, \dots, n+1\}$, that we would like to group by similarity of topic. One strategy is to spot certain words $j \in \{1, \dots, p\}$ in these documents, and compute the frequency f_{ij} – this can be automatized efficiently – of occurrence of the word j in document i. We then form the $n \times p$ matrix X whose (i, j)-th entry is:

$$x_{ij} = f_{ij} - \frac{1}{n+1} \sum_{k=1}^{n+1} f_{kj} \,. \tag{1}$$

Since we subtracted the mean frequency, the data $x_{n+1,j} = -\frac{1}{n+1} \sum_{k=1}^{n} f_{kj}$ is determined by the x_{ij} for $i \leq n$, so it is enough to consider a $n \times p$ matrix.



Let us consider the covariance matrix $M = p^{-1} X X^T$ (X^T is the transpose of the matrix X). M is a symmetric matrix of size $n \times n$, with entries:

$$M_{ik} = \frac{1}{p} \sum_{j=1}^{p} x_{ij} x_{kj} \,.$$

M_{ik} is large when, there are many words $j \in \{1, \dots, p\}$ whose frequency is above the mean both in document i and k, or below the mean both in i and k.

So, M_{ik} can be considered as a measure of the correlation between the documents. For instance, if two documents both contain many "horse" and "ger", but very few "kangaroo" and "bush", the corresponding entry in the matrix M will at least be made of 4 large positive terms. On the other hand, there might be many words – for instance "river", "road", "car", "bird" – whose frequency is close to what can be expected in an arbitrarily chosen document (clearly, one should not choose such generic words, unless one expects them for some reason to be able to differentiate the documents one wants to analyze) ; and some other words – "tea", "cheese", "mountain" – may sometimes appear in excess, or not very frequently, so that the sign of $x_{ij}x_{kj}$ is sometimes positive and negative without a clear trend : in these two cases, the total contribution of these words to M_{ik} will be small in absolute value.

Instead of trying to group documents one by one when we notice a strong correlation – as one can read from the large matrix M – one introduces the notion of **weighted document**, i.e. the assignment of real numbers w_i to each document i. They can be collected in a column vector $W = (w_i)_{1\leq i\leq n}$. Actually, only the relative weight of i and j matters: for any $\lambda > 0$, W and λW represent the same weighted document. A way to fix this ambiguity is to restrict ourselves to vectors W with unit euclidean norm:

$$W^T W = \sum_{i=1}^{n} w_i^2 = 1\,.$$

Then, only W and $-W$ represent the same weighted document. Let us try to find the weighted document W that would display the strongest correlation, i.e. we want to maximize:

$$W^T M W = \sum_{i,k=1}^{n} w_i w_k\, M_{ik}\,.$$

among vectors of unit norm. The answer is that W should be an eigenvector[1] of M with maximum eigenvalue:

$$MW^{(1)} = \lambda_1 W^{(1)}\,.$$

If $W_i^{(1)}$ and $W_j^{(1)}$ are both large and positive – or both large and negative – we can interpret documents i and j as being "similar" according to the strongest pattern that has been found in the data. If $W_i^{(1)}$ is close to 0, it means that the document i does not really participate to this strongest pattern.

We could also have a look at the second, the third, etc. strongest patterns, i.e. consider the eigenvectors $W^{(a)}$ for the a-th eigenvalue, sorted in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Unless the matrix M enjoys for a special reason extra symmetries on top of $M^T = M$, the n eigenvalues computed from the numerical data of M will most likely be distinct, so there is for each $\lambda^{(a)}$ a unique (up to overall sign) eigenvector $W^{(a)}$. Let $E_a = \mathrm{span}(W^{(a)})$ be the eigenspace for λ_a. This method provides a decomposition of the space of

[1] Remember that a symmetric matrix of size $n \times n$ with real-valued entries has exactly n real eigenvalues, counted with multiplicity. In particular, there is a maximum eigenvalue.

weighted documents $\mathbb{R}^n$ into subspaces E_1, $E_1 \oplus E_2$, $E_1 \oplus E_2 \oplus E_3$, ... of dimension $1, 2, 3, \ldots$ In other words, it achieves the task of identifying some low dimensional subspaces $\bigoplus_{\lambda > x} E_\lambda$ in the high dimensional $\mathbb{R}^n$, and the threshold x and dimension gives an indication of the relevance of the pattern that are identified in this way. This method is called **Principal Component Analysis (PCA)**, and was introduced in statistics by Pearson in 1901 [27] and Hotelling in 1931 [20].

To present PCA results, it is customary to draw in the 2-dimensional plane a point with coordinates $p_i = (x_i, y_i)$ with coordinates $x_i = W_i^{(1)}$ and $y_i = W_i^{(2)}$ for each $i \in \{1, \ldots, n\}$. The documents that appear in the same region are then interpreted as "similar" (see Figure 1).

1.2 *Population genetics*

If one replaces "document" by "individual", and "word" by allele (i.e. version) of a gene, the same strategy allows to study the genetic proximity of various populations, and maybe gain some insight into the history of population mixtures. Figure 1 is drawn from such an example.

1.3 *A remark*

From the matrix X, one could also build a $p \times p$ covariance matrix, whose lines and columns are indexed by words (or genes):

$$\tilde{M} = n^{-1} X^T X .$$

Its PCA analysis is useful for factor analysis, i.e. to study what are the most prominent reasons of similarity among the documents (or individuals).

1.4 *A word of caution*

As in any statistical analysis, care should be taken before drawing any conclusion of a cloud of points. PCA has a wide scope of applications in various disciplines, and as a result of its popularity, some research works which use PCA are not free of basic methodology errors. For instance, the most obvious fact is that points gathered near $(0, 0)$ do not represent any information, except that the patterns identified do not allow to distinguish those documents. Another common mistake is to display, say $W^{(3)}$ in abscissa and (to exaggerate) $W^{(18)}$, without questioning the relevance of the eigenvector for the 18-th eigenvalue. It is totally possible that a very small number – like 0, 1, 2, ... – of eigenvectors are actually relevant, the other being not distinguishable from those of a matrix with random entries.

1.5 *The use of random matrix theory*

Random matrix theory provides statistical tests for the relevance of PCA results, as follows. One chooses a **null model**, which in the previous examples would be an ensemble of symmetric random matrices M^{null}. The idea behind

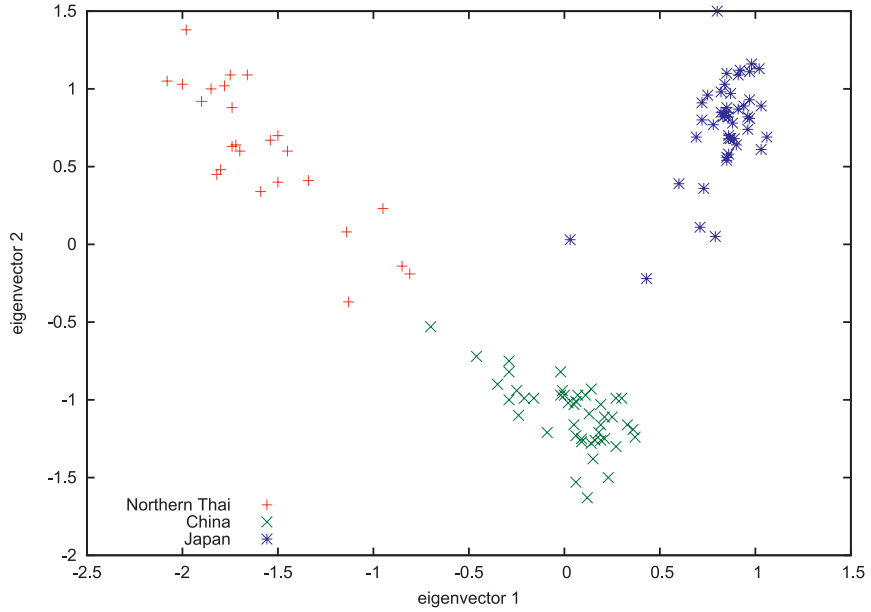


Figure 1: PCA analysis of genetic data of individuals from 3 East Asian populations, based on the International Haplotype Map, and concerning $p = 40560$ SNPs. SNP stands for Single Nucleotide Polymorphism: genes come in several versions, which often differ by the nature of the nucleotide (A, C, G or T) present in a few specific positions in the gene. Up to a correction factor, $f_{i,j}$ in (1) measures the frequency of a given allele (=version of a gene) j carried by an individual i, and therefore takes values 0, 1 or 2 (this last case means that the two chromosomes carry the same allele). Reprinted from *Population structure and eigenanalysis*, N. Patterson, A.L. Preis and D. Reich, PLoS Genetics **2** 12 (2006).

the choice of the null model is that sampling M^{null} in this random ensemble will produce data that "contain no information" compared to the type of information we would like to identify in genuine data. Imagine that one has computed the probability p_A^{null} of various events A concerning the eigenvalues or the eigenvectors of a matrix M drawn from the null model. If one observes the event A in the genuine data one is analyzing, we say that the null model can be rejected with confidence $1 - p_A^{\text{null}}$.

To this end, for various random ensembles of matrices (that one could take as null models):

- we need to know the distribution of eigenvalues, especially in the limit of matrices of large size ;
- we are especially interested in extreme (maximal or minimal) eigenvalues ;

- and we would like to understand whether these distributions are very sensitive or not to the choice of the null model, i.e. what happens to the spectrum if we do small perturbations of our random matrix.

These questions are a priori non obvious to answer, and represent typical interests in random matrix theory.

2 General principles

We shall introduce in Section 3 and 4 two ensembles of random matrices, but before that, let us pose the problem in mathematical terms.

2.1 *Definition and tools*

We say that a $n \times n$ matrix M is symmetric if M_{ij} is real and $M_{ij} = M_{ji}$, and that is hermitian is M_{ij} is complex and $M_{ij} = M_{ji}^*$ where the * stands for complex conjugate. We denote:

$$\mathcal{S}_n = \{n \times n \text{ symmetric matrices}\}, \qquad \mathcal{H}_n = \{n \times n \text{ hermitian matrices}\} \tag{2}$$

and we note that $\mathcal{S}_n \subseteq \mathcal{H}_n$. The Lebesgue measure on $\mathcal{S}_n$ is by definition the product of the Lebesgue measures on the linearly independent entries of M:

$$\mathrm{d}M = \prod_{1 \le i < j \le n} \mathrm{d}M_{ij} \prod_{i=1}^{n} \mathrm{d}M_{ii} \,.$$

Similarly on $\mathcal{H}_n$:

$$\mathrm{d}M = \prod_{1 \le i < j \le n} \mathrm{d}(\operatorname{Re} M_{ij})\, \mathrm{d}(\operatorname{Im} M_{ij}) \prod_{i=1}^{n} \mathrm{d}M_{ii} \,.$$

A matrix $M \in \mathcal{H}_n$ has exactly n real eigenvalues, that we write in decreasing order:

$$\lambda_1^{(M)} \ge \lambda_2^{(M)} \ge \cdots \ge \lambda_n^{(M)} \,.$$

The spectral measure is the probability measure:

$$L^{(M)} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i^{(M)}} \,.$$

consisting of a Dirac mass $1/n$ on each eigenvalue. This is a convenient way to collect information on the spectrum of M, since for any continuous function f, we can write:

$$\sum_{i=1}^{n} f(\lambda_i^{(M)}) = \int f(x)\, \mathrm{d}L^{(M)}(x) \,.$$

We state without proof the **Hoffman-Wielandt inequality**:

$$\forall A, B \in \mathcal{H}_n, \qquad \sum_{i=1}^{n} (\lambda_i^A - \lambda_i^B)^2 \le \operatorname{Tr} (A - B)^2 \,.$$

The right-hand side can be written in several forms:

$$\operatorname{Tr} M^2 = \sum_{i=1}^{n} (\lambda_i^{(M)})^2 = \sum_{i,j=1}^{n} M_{ij} M_{ji} = \sum_{i,j=1}^{n} |M_{ij}|^2 \,.$$

We remark that, since A and B a priori do not commute, $\lambda_i^A - \lambda_i^B$ is not in general an eigenvalue of $A - B$. This inequality is pretty useful. For instance, it tells us that the vector of eigenvalues $(\lambda_1^{(M)}, \dots, \lambda_n^{(M)})$ is a Lipschitz – and a fortiori, continuous[2] – function of the entries of M.

2.2 *Random matrices, topology, convergence*

By convention, any topological space is equipped with the σ-algebra generated by its open sets – the so-called Borel σ-algebra.

A random matrix of size n is a random variable M_n with values in $\mathcal{H}_n$, i.e. a measurable function from a set Ω to $\mathcal{H}_n$. Since eigenvalues are continuous functions of the entries, the $\lambda_i^{(M_n)}$ are also random variables, i.e. measurable functions from Ω to $\mathbb{R}$. The random probability measure $L^{(M_n)}$ is called the **empirical (spectral) measure**. At this point we need to specify the topology we choose on the set $\mathcal{M}^1(\mathbb{R})$ of probability measures on $\mathbb{R}$. We shall be concerned with two choices: the **weak topology** and the **vague topology**. For the weak topology, $\mathcal{M}^1(\mathbb{R})$ is a Polish space ; as a consequence (or as a fact for those who are not familiar with topology), it is enough to declare what does it mean for a sequence $(\mu_n)_n$ of probability measures to converge to a probability measure μ in this topology:

$$\mu_n \xrightarrow[n\infty]{\text{weak}} \mu \qquad \Longleftrightarrow \qquad \forall f \in \mathcal{C}_b^0, \qquad \lim_{n\to\infty} \int f \mathrm{d}\mu_n = \int f \mathrm{d}\mu \,,$$

where $\mathcal{C}_b^0$ is the set of continuous bounded functions from $\mathbb{R}$ to $\mathbb{R}$. For the vague topology, the convergence of sequences is nearly the same:

$$\mu_n \xrightarrow[n\infty]{\text{vague}} \mu \qquad \Longleftrightarrow \qquad \forall f \in \mathcal{C}_c^0, \qquad \lim_{n\to\infty} \int f \mathrm{d}\mu_n = \int f \mathrm{d}\mu \,,$$

where $\mathcal{C}_c^0$ is the set of continuous functions with compact support. Therefore, convergence for the weak topology implies convergence for the vague topology, but the converse may not hold. Now, if we equip $\mathcal{M}^1(\mathbb{R})$ is equipped with the Borel σ-algebra of any of these topologies, the empirical measure $L^{(M_n)}$ is a (probability measure)-valued random variable, i.e. a measurable function $\Omega \to \mathcal{M}^1(\mathbb{R})$.

Usually, we are dealing with a ensemble of random matrices for each n, and want to study the spectrum when $n \to \infty$. We should distinguish:

- **global information**, which involve the macroscopic behavior of eigenvalues. For instance, we ask about the convergence of $L^{(M_n)}$ – as a random variable – towards a deterministic limit, its fluctuations, etc.
- and **local information**, which concern only $O(1)$ eigenvalues. For instance, we ask about the convergence of the maximal eigenvalue $\lambda_1^{(M_n)}$,

[2] Another way to prove this is to remark that the eigenvalues of M are the roots of the characteristic polynomial $\det(z - M)$. The coefficients of this polynomial of z are polynomial functions of the entries of M, thus continuous, and it is a standard result of complex analysis that the roots of a polynomial are continuous functions of the coefficients.

its fluctuations, etc.

We remind that, if $(X_n)_n$ is a sequence of random variables with values in $\mathcal{X}$, there are several (non-equivalent) notions of convergence to another $\mathcal{X}$-valued random variable X. The three main ones we shall use are **almost sure** convergence, convergence in **probability** and for $\mathcal{X} = \mathbb{R}$, convergence **in law**. The definitions are "$(X_n)_n$ converges to $X \dots$"

- almost surely, if $\mathbb{P}\left[\lim_{n\to\infty} X_n = X\right] = 1$.
- in probability, if for any $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left[|X_n - X| > \epsilon\right] = 0$.
- in law, if for any $x \in \mathbb{R}$ at which $\mathbb{P}[X \leq x]$ is continuous,

$$\lim_{n\to\infty} \mathbb{P}[X_n \leq x] = \mathbb{P}[X \leq x].$$

We remind that almost sure convergence implies convergence in probability, and the latter implies convergence in law, but the converse in general do not hold.

Even if the entries M_{ij} are independent random variables, the eigenvalues depend in a non-linear way of all the entries, and therefore are strongly correlated. For this reason, the limit distributions of the spectrum in the limit $n \to \infty$ are in general very different than the limit distributions one can find in the theory of independent random variables[3]. We will see a few of these new limit laws in the lectures. It turns out these laws enjoy some universality, and the results of random matrix theory have found applications way beyond statistics, e.g. in biology and the study of ARN folding, in number theory, in nuclear physics, statistical physics and string theory, etc.

2.3 *Qualitative remarks*

Size of the spectrum

Imagine that one fills a hermitian matrix M_n of size n with entries of size $O(1)$. How large (as a function of n) in absolute value can we expect the eigenvalues to be ? We have:

$$\operatorname{Tr} M_n^2 = \sum_{i,j=1}^{n} \left|[M_n]_{ij}\right|^2 = \sum_{i=1}^{n} \left[\lambda_i^{(M_n)}\right]^2 .$$

This quantity is of order n^2, since in the first expression it is written as a sum of n^2 terms of order 1. Then, from the second expression we deduce roughly that the eigenvalues should be order $\sqrt{n}$. In other words, if we fill a matrix M_n of size n with entries of size $O(n^{-1/2})$ – or equivalently with random variables having variance of order of magnitude $1/n$ – we can expect

[3] For independent identically distributed random variables, we have the law of large numbers and the central limit theorem for the sum, and we also know that the possible limit distributions for the maximum of a sequence of i.i.d. are the Gumbel law (e.g. for variables whose distribution decays exponentially), the Fréchet law (e.g. for heavy tailed distributions) and the Weibull law (e.g. for bounded random variables).

the spectrum to remain bounded when $n \to \infty$. This non-rigorous argument serves as an explanation of the scalings in forthcoming definitions.

Stability under perturbations

Let M_n be a random matrix of size n, and assume that when $n \to \infty$, $L^{(M_n)}$ converges to a deterministic limit μ in probability for the vague topology, i.e. for any $\epsilon > 0$ and $f \in \mathcal{C}_c^0$,

$$\lim_{n\to\infty} \mathbb{P}\Big[\Big|\int f(x)\,\mathrm{d}(\mu_n - \mu)(x)\Big| > \epsilon\Big] = 0\,. \tag{3}$$

Then, let Δ_n be another random matrix of size n.

2.1 lemma. *If $\lim_{n\to\infty} n^{-1}\mathbb{E}[\operatorname{Tr}\Delta_n^2] = 0$, then $L^{(M_n+\Delta_n)}$ converges to μ in probability, for the vague topology.*

Proof. Any continuous f with compact support can be approximated for the sup norm by a polynomial (Stone-Weierstraß theorem), in particular by a Lipschitz function. Therefore, it is enough to prove that (3) holds for $\mu_n = L^{(M_n+\Delta_n)}$ for any $\epsilon > 0$ and f Lipschitz. Let us denote k its Lipschitz constant. We have:

$$\begin{aligned}
\Big|\int f(x)\mathrm{d}(L^{(M_n+\Delta_n)} - \mathrm{d}L^{(M_n)})(x)\Big| &= \frac{1}{n}\Big|\sum_{i=1}^n f(\lambda_i^{(M_n+\Delta_n)}) - f(\lambda_i^{(M_n)})\Big| \\
&\le \frac{1}{n}\sum_{i=1}^n k\big|\lambda_i^{(M_n+\Delta_n)} - \lambda_i^{(M_n)}\big| \\
&\le \frac{k}{\sqrt{n}}\Big(\sum_{i=1}^N (\lambda_i^{(M_n+\Delta_n)} - \lambda_i^{(M_n)})^2\Big)^{1/2} \\
&\le \frac{k}{\sqrt{n}}\big(\operatorname{Tr}\Delta_n^2\big)^{1/2},
\end{aligned}$$

where we have used Cauchy-Schwarz inequality, and the Hoffman-Wielandt inequality. Then, for any fixed $\epsilon > 0$, with Markov inequality:

$$\mathbb{P}\Big[\Big|\int f(x)\,\mathrm{d}(L^{(M_n+\Delta_n)} - L^{(M_n)})(x)\Big| > \epsilon\Big] \le \frac{k^2\,\mathbb{E}[\operatorname{Tr}\Delta_n^2]}{n\epsilon^2}\,,$$

and under the assumption of the lemma, the right-hand side converges to 0. Since we already had (3) for $\mu_n = L^{(M_n)}$, we have proved the desired result. □

As we have seen before, it is natural to consider matrices M_n whose entries have variance bounded by C/n. In that case, according to this lemma, we could make $o(n^2)$ entries deterministic – by choosing $[\Delta_n]_{ij} = \mathbb{E}[[M_n]_{ij}] - [M_n]_{ij}$ for the selected entries – without affecting the convergence of the empirical measure to the limit μ. This lemma indicates that small perturbations of a random matrix do not affect global properties of the spectrum.

There is no such general rule for local properties (such as the position of the maximum eigenvalue) : we will see examples showing that sometimes they

are preserved under small perturbations, and sometimes they are dramatically affected.

3 Wishart matrices

3.1 *Definition*

A **real Wishart matrix** is a random symmetric matrix M of the form:

$$M = n^{-1} X^T X ,$$

where X is random matrix of size $n \times p$ such that:

- $(X_{ij})_{1 \leq i \leq n}$ are independent samples of a real-valued random variable $\mathcal{X}_j$;
- $(\mathcal{X}_1, \dots, \mathcal{X}_p)$ is a Gaussian vector with given covariance $K \in \mathcal{S}_p$

In other words, the joint probability density function (= p.d.f.) of the entries of X is:

$$c_{np}(K) \exp\Big(-\frac{1}{2} \sum_{i,i'=1}^{n} \sum_{j,j'=1}^{p} X_{ij} X_{i'j'} K^{-1}_{jj'} \Big) = c_{np}(K) \exp\Big(-\frac{1}{2} \mathrm{Tr}\, X^T K^{-1} X \Big) .$$

$c_{np}(K)$ is a normalization constant. All the normalization constants that will appear in these lectures can be explicitly computed, but we will not care about them. The matrix M is of size $p \times p$, and n is called the number of degrees of freedom. The parameter:

$$\gamma = n/p$$

will play an important role. The ensemble of real Wishart matrices with a covariance $K = \mathrm{diag}(\sigma^2, \dots, \sigma^2)$ is a natural choice of null model for covariance matrices in data analysis, which depends on a parameter σ. It was introduced by Wishart in 1928 [37].

One can also define the ensemble of **complex Wishart matrices**. These are random hermitian matrices of the form $M = (X^T)^* X$, where $(X_{ij})_{1 \leq i \leq n}$ are independent samples of $\mathcal{X}_j$ such that $(\mathcal{X}_1, \dots, \mathcal{X}_p)$ is a complex Gaussian vector with given covariance $K \in \mathcal{H}_p$. This is one of the simplest model of complex random matrices, and the latter are relevant e.g. in telecommunications, when one studies non-ideal propagation of waves along many canals (complex numbers are used to encode simultaneously the amplitude and the phase of a wave).

3.2 *Spectral density in the large size limit*

We consider real or complex Wishart ensembles with given covariance $K = \mathrm{diag}(\sigma^2, \dots, \sigma^2)$. Marčenko and Pastur showed in 1967 [25] that the empirical measure $L^{(M)}$ has a deterministic limit:

3.1 theorem. *If the limit where $p, n \to \infty$ while n/p converges to a fixed value $\gamma \in (0, +\infty)$, $L^{(M)}$ converges almost surely and in expectation in the weak topology, towards the probability measure (see Figure 2):*

$$\mu_{\mathrm{MP}} = \max(1 - \gamma, 0)\delta_0 + \frac{\gamma \sqrt{(a_+(\gamma) - x)(x - a_-(\gamma))}}{2\pi\sigma^2 \, x} \mathbf{1}_{[a_-(\gamma), a_+(\gamma)]} \, \mathrm{d}x \tag{4}$$

where $a_\pm(\gamma) = \sigma^2(1 \pm \gamma^{-1/2})^2$.

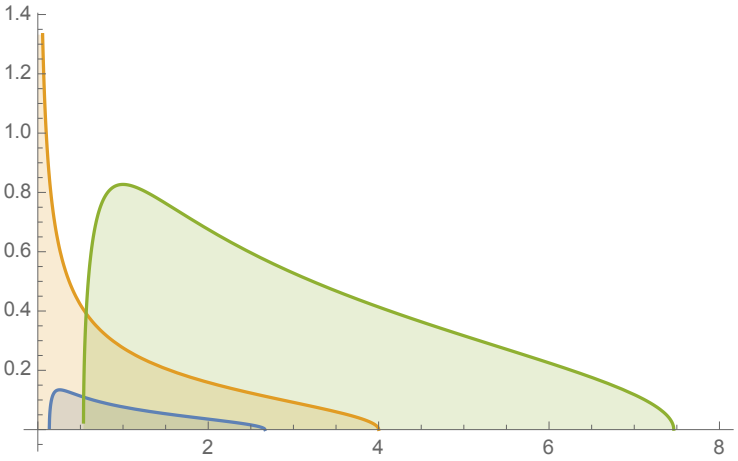


Figure 2: Marčenko-Pastur probability density function, for $\sigma^2 = 1$: in green $\gamma = 3$, in orange $\gamma = 1$, in blue $\gamma = 0.4$. The mass of the distribution in this last case is 0.4, to which should be added a Dirac mass with mass 0.6 at 0.

We note that when $n < p$, the matrix $X^T X$ has rank $n < p$, and therefore has almost surely $p - n = p(1-\gamma)$ zero eigenvalues, which explains the Dirac mass in (4) which appear for $\gamma < 1$. The mean and variance of the Marčenko-Pastur distribution are:

$$\int x \, \mathrm{d}\mu_{\mathrm{MP}}(x) = \sigma^2, \qquad \int x^2 \, \mathrm{d}\mu_{\mathrm{MP}}(x) - \Big(\int x \mathrm{d}\mu_{\mathrm{MP}}(x) \Big)^2 = \sigma^4/\gamma \,. \tag{5}$$

Apart from the possible Dirac mass at 0, the support of μ_{MP} is spread on an interval of length $4\sigma^2\gamma^{-1/2}$ around the mean σ^2: the smaller γ is, the broader the support becomes. On the other hand, when $\gamma \to \infty$, the support becomes localized around σ^2, i.e. we can read the variance of the Gaussian entries of X. For practical applications, this means that if the number of measurements n is not very large compared to the number p of properties we measure, the spectrum of M will be spread.

Another property of μ_{MP} is that, for[4] $\gamma \neq 1$, the density of μ_{MP} vanishes like a squareroot at the edges $a_\pm(\gamma)$. This behavior is frequent for the spectra of large random matrices.

3.3 *Maximum eigenvalue and fluctuations*

From Marčenko-Pastur theorem, one can easily deduce that, for any $\epsilon > 0$,

$$\mathbb{P}[\lambda_1^{(M)} \leq a_+(\gamma) - \epsilon] \to 0 \,,$$

[4] For $\gamma = 1$, it diverges as $x^{-1/2}$ when $x \to 0^+$.

and thus that $(\limsup_{n\to\infty} \lambda_1^{(M)})$ is almost surely larger than $a_+(\gamma)$. Indeed, let us choose an arbitrary non-negative, non-zero, continuous function f with compact support included in $(a_+(\gamma)-\epsilon, +\infty)$. We can rescale f to enforce $\int f(x)\mathrm{d}\mu_{\mathrm{MP}}(x) = 1$. We then have:

$$\begin{aligned}\mathbb{P}\left[\lambda_1^{(M)} \leq a_+(\gamma)-\epsilon\right] &\leq \mathbb{P}\Big[\int f(x)\,\mathrm{d}L^{(M)}(x) = 0\Big] \\ &\leq \mathbb{P}\Big[\Big|\int f(x)\mathrm{d}(L^{(M)}-\mu_{\mathrm{MP}})(x)\Big| \geq 1/2\Big]\,,\end{aligned}$$

and the latter converges to 0 when $n, p \to \infty$ according to Theorem 3.1. But Theorem 3.1 does not tell us whether the maximum eigenvalue $\lambda_1^{(M)}$ really converges to $a_+(\gamma)$ or not. The reason is easily understood: the event $\lambda_1^{(M)} \leq a_+(\gamma)-\epsilon$ actually means that all eigenvalues are smaller than $a_+(\gamma)-\epsilon$: this is a global information, hence contained in the statement of convergence of $L^{(M)}$. However, the realization of an event like $\lambda_1^{(M)} \geq a_+(\gamma)-\epsilon$ only involves a single eigenvalue, and thus more work is needed to estimate its probability. We will not say how this work is done, but the result is that there is no surprise:

3.2 theorem. *[16] $\lambda_1^{(M)}$ converges almost surely to $a_+(\gamma)$.*

The distribution of the fluctuations of $\lambda_1^{(M)}$ is also known. Before presenting the result, let us give a non-rigorous argument to guess the order of magnitude of these fluctuations. The guess is that, for a Wishart matrix of large size p, the number of eigenvalues in an interval I_p whose length depend on p should be well approximated by $p\mu_{\mathrm{MP}}[I_p]$. So, we guess that the fluctuations of $\lambda_1^{(M)}$ should occur in a region of width $\delta_p \to 0$ around $a_+(\gamma)$ where μ_{MP} has mass of order $1/p$. Since μ_{MP} vanishes like a squareroot at the edge, we have:

$$\mu_{\mathrm{MP}}[a_+(\gamma)-\delta_p, a_+(\gamma)] \sim \int_0^{\delta_p} x^{1/2}\mathrm{d}x = \frac{2}{3}\delta_p^{3/2}\,,$$

and this gives the estimate $\delta_p \sim p^{-2/3}$. The following result [15, 22] confirms this guess:

3.3 theorem. *We set $\beta = 1$ for real Wishart, and $\beta = 2$ for complex Wishart. The random variable:*

$$\gamma^{1/2}p^{2/3}\,\frac{\lambda_1^{(M)} - a_+(\gamma)}{\sigma^2(1+\gamma^{-1/2})^{4/3}}$$

converges in law towards a random variable Ξ_β when $n, p \to \infty$ while n/p converges to $\gamma \in (0, +\infty)$.

The distribution function:

$$\mathrm{TW}_\beta(s) = \mathbb{P}[\Xi_\beta \leq s]$$

is called the **Tracy-Widom law**. It is not an elementary function, but can be considered as a new special function. It is nowadays well-tabulated, hence ready for use in statistics (Figure 3). We now give one of their expression, first

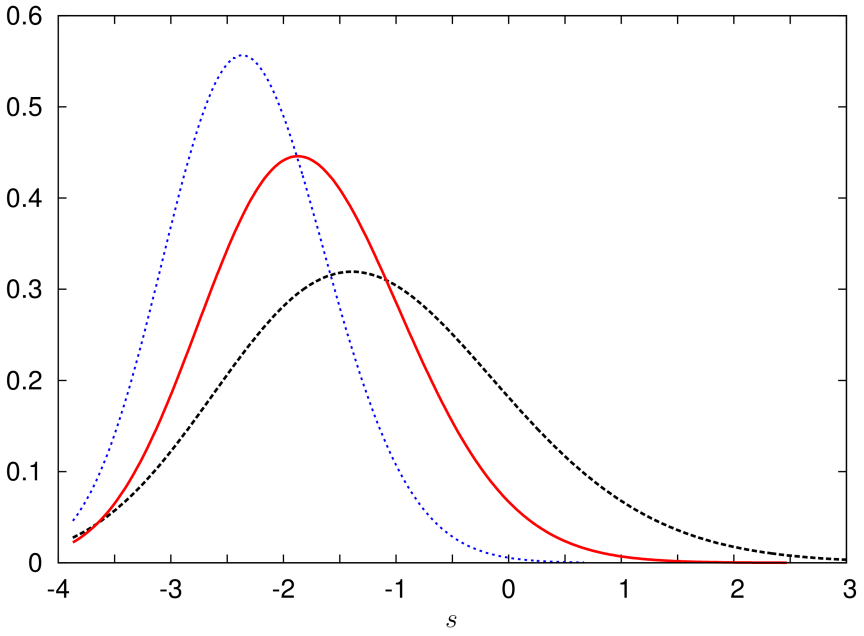


Figure 3: Probability density function of the Tracy-Widom law, i.e. $\mathrm{TW}_\beta(s)$, for $\beta = 1$ (GOE, in blue), $\beta = 2$ (GUE, in red), and $\beta = 4$. Graph courtesy of J.M. Stéphan.

obtained by Tracy and Widom in 1992 for $\beta = 2$ [33] and 1995 for $\beta = 1$ [34]:

$$\begin{aligned} \text{hermitian} \qquad \mathrm{TW}_2(s) &= \exp\Big[-\int_s^\infty \{q'(t) - tq^2(t) - q^4(t)\}\mathrm{d}t\Big], \\ \text{symmetric} \qquad \mathrm{TW}_1(s) &= \exp\Big[-\frac{1}{2}\int_s^\infty q(t)\mathrm{d}t\Big]. \end{aligned}$$

Here, $q(t)$ is the unique bounded solution to the **Painlevé II equation**:

$$q''(t) = 2q^3(t) + tq(t)$$

satisfying the growth conditions $q(t) \sim \sqrt{-t/2}$ when $t \to -\infty$, and:

$$q(t) \sim \frac{\exp(-\frac{2}{3}t^{3/2})}{2\sqrt{\pi}t^{1/4}}, \qquad t \to +\infty\,.$$

Existence and uniqueness of the function $q(t)$ was shown by Hastings and McLeod in 1980 [19], and it bears their name. We will derive in Section 9.3 an-

other expression for $\mathrm{TW}_2(s)$ in terms of a infinite size (Fredholm) determinant, which is actually the easiest way to compute numerically the Tracy-Widom law.

3.4 *Application to Markowitz portfolio optimization*

This paragraph is based on the article *Random matrix theory and financial correlations*, Bouchaud, Cizeau, Laloux, Potters, Risk Magazine **12** 69 (1999), and the figures extracted from this article.

Imagine we consider investing in assets $j \in \{1, \dots, p\}$ a fraction of money w_j. We would like to determine, for a fixed return r, to determine the choice of portfolio $(w_1^*, \dots, w_n^*)$ minimizing the risk. For this purpose, we only have at our disposal the observations of the price p_{ij} of these assets at times $i \in \{1, \dots, n\}$ in the past. We can subtract the mean price and write $p_{ij} = \overline{p}_j + x_{ij}$. If we had invested in the past and get our return at time i, we would have earned:

$$r_i = \sum_{j=1}^{p} w_j(\overline{p}_j + x_{ij})$$

If we are ready to believe[5] that these observations represent well of what can happen during the (future) period of our investment, we can take:

$$r = \sum_{j=1}^{p} w_j \overline{p}_j + \frac{1}{n} \sum_{i=1}^{n} w_j x_{ij} = \overline{r} + J^T X W$$

where W is column vector representing the portfolio, J the column vector with entries $1/n$, and $X = (x_{ij})_{ij}$ the $n \times p$ matrix collecting the observations. One can also try to evaluate the risk in investing as W with the quantity:

$$\rho = \sum_{j,j'=1}^{p} w_j w_{j'} \Big(\frac{1}{n} \sum_{i=1}^{n} x_{ij} x_{ij'} \Big) = W^T M W \,,$$

where:

$$M = n^{-1} X^T X$$

is the empirical correlation matrix. Finding the W^* that minimizes ρ for a given $(r - \overline{r})$ can be done by minimizing the quantity $W^T M W - a J^T X W$ for a constant a – the Lagrange multiplier – that we adjust so that:

$$r - \overline{r}) = W^T M W \,.$$

Denoting $P = J^T X$, the result is:

$$\rho^* = \frac{(r - \overline{r})^2}{P^T M^{-1} P} \qquad W^* = \frac{\rho^*}{r - \overline{r}} M^{-1} P \,. \tag{6}$$

In particular, we see that the eigenvectors of M with small eigenvalues play an important role in the evaluation of ρ^* and W^*. This is the base of the

[5] This is highly criticizable, especially in finance. We will come back to this point.

method proposed by Markowitz in 1952 [24]. One usually plots the return r as a function of the estimation ρ^* of the risk: the curve is called the **efficient frontier**, and in this simple model, it is a parabola.

As a matter of fact, it is hard to build an empirical covariance matrix reliable for future investments, and Markowitz theory suffers in practice from important biases. With an example drawn from genuine financial data, Bouchaud et al. pointed out that a large part – and especially the lower part – of the spectrum of M can be fitted with a Marčenko-Pastur distribution, hence cannot be distinguished from the null model of large random covariance matrix (Figure 4). The effect is that the minimal risk for a given return is underestimated (Figure 6), and the guess (6) of the optimal portfolio does not give good results.

The part of the spectrum undistinguishable from noise is called the **noise band**. If one make observations of the prices and build empirical correlation matrices over two distinct periods, one can also check that the eigenvectors for eigenvalues outside the noise band have common features – quantitatively measured by the absolute value of their scalar product – while the eigenvectors for eigenvalues in the noise band have nothing more in common than two random vectors (Figure 5). It supports the idea that only eigenvectors for eigenvalues outside the noise band contain a genuine information about the long-time evolution of the market.

Although there is no ideal cure, Bouchaud et al. proposed to replace the empirical correlation matrix M by $\widetilde{M}$ build as follows.

- Decompose $\mathbb{R}^n = E_{\text{noise}} \oplus E$, where E_{noise} (resp. E) is the sum of eigenspaces for eigenvalues in the noise band (resp. outside the noise band).
- Replace the restriction of M to E_{noise} by a multiple of the identity operator, so that the trace is preserved.
- Use the new matrix $\widetilde{M}$ in the Markowitz optimization formulas (6).

The risk is still underestimated, but to a smaller extent.

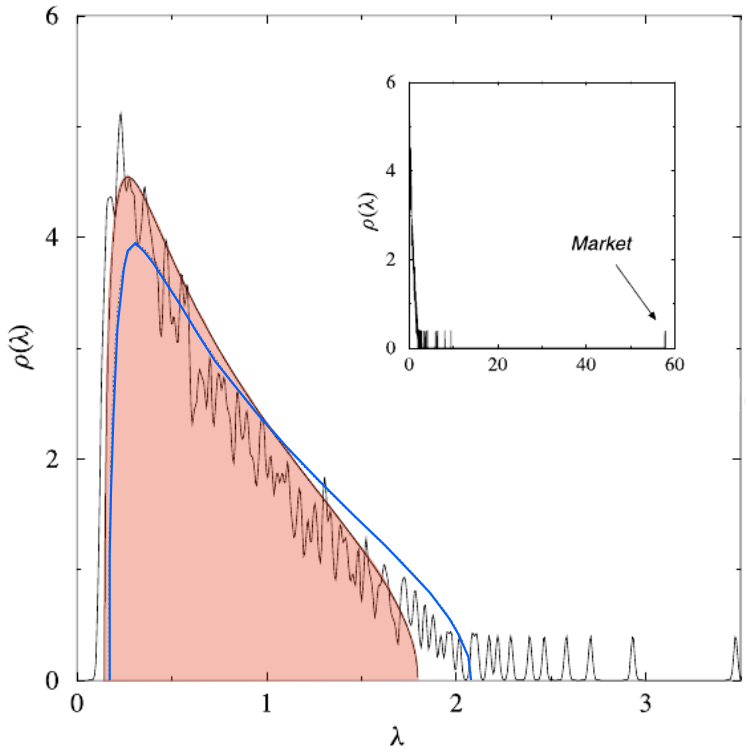


Figure 4: Spectrum of an empirical $p \times p$ covariance matrix, build from the value of $p = 406$ assets from the S&P 500, observed every day in a period of $n = 1309$ days between 1991 and 1996. One eigenvalue is much larger than the other, and correspond to the market mode, i.e. all assets increase or decrease simultaneously. The blue (resp. red) curve is the Marčenko-Pastur (MP) spectral density for a large Wishart matrix with $\gamma = n/p$, and input covariance $\mathrm{diag}(\sigma^2, \ldots, \sigma^2)$ for $\sigma^2 = 0.85$ (resp. $\sigma^2 = 0.74$). This last value is the optimal fit. About 6% of the eigenvalues cannot be not accounted by the MP law, and they are responsible represent $1 - \sigma^2 = 26\%$ of the variance. We note that the shape of the empirical density of low eigenvalues is well reproduced by MP, so these eigenvalues (and the corresponding eigenvectors, which have the largest weight for Markowitz optimization) cannot be distinguished from noise.

Lecture 2 (1h30)
August 6th, 2015

4 Gaussian ensembles

The Gaussian ensembles are the simplest ensembles of random matrices from the computational point of view. As Wishart matrices, they come in two flavors, depending whether one considers symmetric or hermitian matrices. For

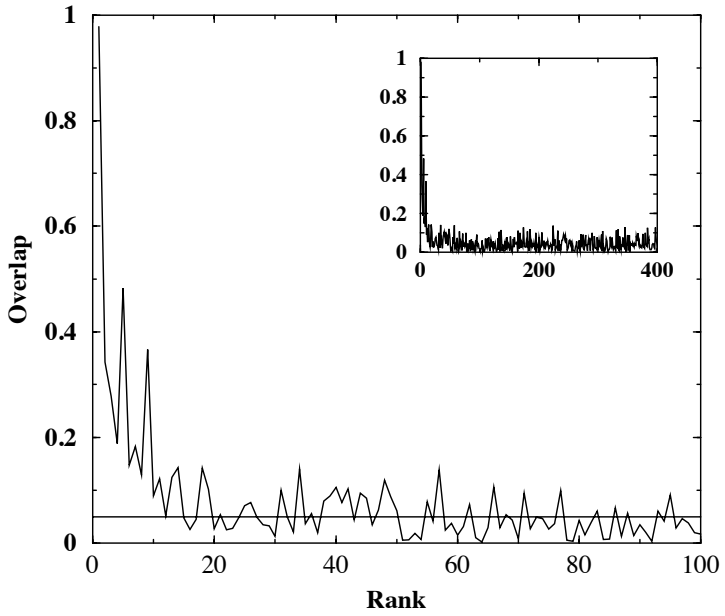


Figure 5: $M^{(1)}$ and $M^{(2)}$ are empirical correlation matrices coming from observation in period 1 and 2. If we denote $W_i^{(a)}$ a unit norm eigenvector of $M^{(a)}$ for the i-th eigenvalue (in decreasing order) with norm 1, the plot shows the scalar product $|W_i^{(1)} \cdot W_i^{(2)}|$ as a function of $i = 1, 2, 3, \ldots$ in abscissa. The horizontal line $1/\sqrt{p}$ is the typical value for the overlap of two independent random vectors with normal entries Gaussian entries.

a reason revealed in Section 7.1, the symmetric case is labeled $\beta = 1$, and the hermitian case $\beta = 2$.

In the **Gaussian Orthogonal Ensemble** (GOE), we consider a symmetric random matrix M of size $n \times n$, with

$$M_{ij} = \begin{cases} X_{ij} & 1 \leq i < j \leq n \\ X_{ji} & 1 \leq j < i \leq n \\ Y_i & 1 \leq i = j \leq n \end{cases}, \tag{7}$$

where X_{ij} and Y_i are independent centered Gaussian random variables with:

$$\mathbb{E}[X_{ij}^2] = \sigma^2/n, \qquad \mathbb{E}[Y_i^2] = 2\sigma^2/n\,. \tag{8}$$

We choose to scale the variance by $1/n$, so that the spectrum will remain bounded – see Section 2.3. The difference of normalization between the off-diagonal and diagonal elements is motivated by observing that the resulting

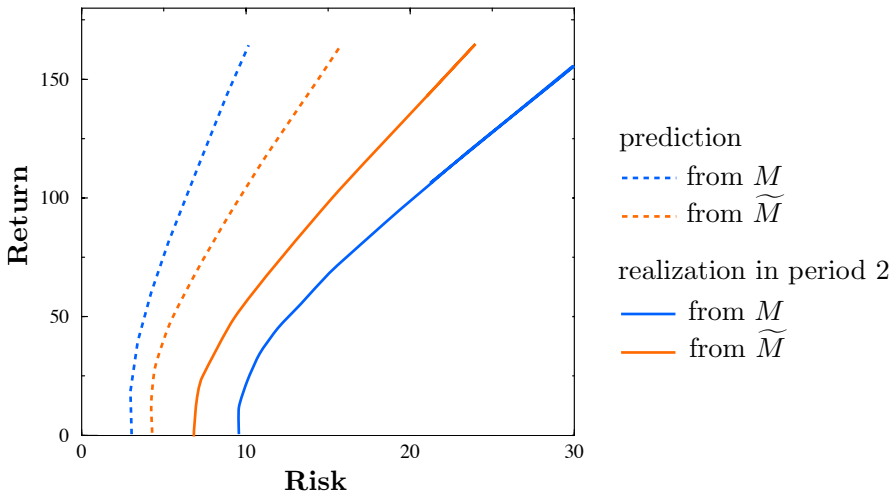


Figure 6: The dashed curve is the prediction from M (blue) or $\widetilde{M}$ (orange) of the effective frontier via (6), constructed from the observations in a first period of time, and proposing to invest W^* or $\widetilde{W}^*$. The plain curves correspond to the effective frontier measured if we really invested W^* (blue) or $\tilde{W}^*$ (orange) in the second period of time.

probability measure on the entries of M is proportional to:

$$(9) \quad \mathrm{d}M \exp\Big[-\frac{n}{2\sigma^2}\Big(2\sum_{i=1}^{n} M_{ii}^2 + \sum_{1\le i<j\le n} M_{ij}^2\Big)\Big] = \mathrm{d}M \exp\Big[-\frac{n}{2\sigma^2}\operatorname{Tr} M^2\Big].$$

The Lebesgue measure $\mathrm{d}M$ is invariant under conjugation $M \mapsto \Omega^{-1}M\Omega$ by an orthogonal matrix Ω, and so is $\operatorname{Tr} M^2$. Therefore, for any orthogonal matrix Ω, M drawn from (9) and $\Omega^{-1}M\Omega$ have the same distribution, and this explains the name GOE. This property would not be true if we had chosen the same variance in (7) for the diagonal and off-diagonal entries.

In the **Gaussian Unitary Ensemble** (GUE), we consider a hermitian random matrix M of size $n \times n$, with

$$M_{ij} = \begin{cases} X_{ij} + \sqrt{-1}\,\tilde{X}_{ij} & 1 \le i < j \le n \\ X_{ji} + \sqrt{-1}\,\tilde{X}_{ji} & 1 \le j < i \le n \\ Y_{ii} & 1 \le i = j \le n \end{cases}$$

where X_{ij}, $\tilde{X}_{ij}$ and Y_i are independent centered Gaussian random variables

with:

$$\mathbb{E}[X_{ij}^2] = \mathbb{E}[\tilde{X}_{ij}^2] = \sigma^2/2n, \qquad \mathbb{E}[Y_i^2] = \sigma^2/n\,.$$

The resulting probability measure on the entries of M reads:

$$\mathrm{d}M \, \exp\Big[-\frac{n}{\sigma^2} \operatorname{Tr} M^2 \Big]\,,$$

and it is invariant under conjugation $M \mapsto \Omega^{-1} M \Omega$ by a unitary matrix Ω.

The probability measures for the GOE and the GUE can written in a unified way:

$$\mathrm{d}M \, \exp\Big[-\frac{n\beta}{2\sigma^2} \operatorname{Tr} M^2 \Big]\,.$$

The results that we have seen in the case of Wishart matrices for the spectral density in the large size limit, and the location of the maximum eigenvalue and its fluctuations, have an analog for the Gaussian ensembles. Their proof in the case $\beta = 2$ (GUE) will be sketched in Section 9.

4.1 *Spectral density*

Let M_n be a random matrix in the GOE or the GUE. Wigner showed in 1955 [36] that the empirical measure $L^{(M_n)}$ converges to a deterministic limit – although the almost sure mode of convergence was only obtained later, by large deviation techniques – see e.g. the book [1].

4.1 theorem. *When $n \to \infty$, $L^{(M_n)}$ converges almost surely and in expectation to the probability measure (see Figure 7 for a plot):*

$$\mu_{\mathrm{sc}} = \frac{\sqrt{4\sigma^2 - x^2}}{2\pi\sigma^2} \mathbf{1}_{[-2\sigma, 2\sigma]}(x)\, \mathrm{d}x\,. \tag{10}$$

μ_{sc} is called the semi-circle law, because of the shape of its density when $\sigma = 1$. It is symmetric around 0, and the variance is:

$$\int_{-2\sigma}^{2\sigma} x^2 \, \mathrm{d}\mu_{\mathrm{sc}}(x) = \sigma^2\,.$$

As in the Wishart case, we observe that the density of μ_{sc} vanishes like a squareroot at the edges of its support.

4.2 *Maximum eigenvalue and fluctuations*

4.2 theorem. *[26] When $n \to \infty$, $\lambda_1^{(M_n)}$ converges almost surely to 2σ. Besides, we have the convergence in law:*

$$n^{2/3}\sigma^{-1}\big\{\lambda_1^{(M_n)} - 2\sigma\big\} \underset{n\infty}{\longrightarrow} \Xi_\beta\,,$$

where Ξ_β is drawn from the Tracy-Widom law with $\beta = 1$ for GOE, and $\beta = 2$ for GUE.

Comparing to the Wishart case, we remark that the global properties of the spectrum do not depend on the type – $\beta = 1$ for symmetric, or $\beta = 2$ for hermitian – of matrices once the ensemble is properly normalized, while the local properties (e.g. the Tracy-Widom laws) depend non-trivially on β, as one can see in Figure 3.

5 Stieltjes transform and freeness

5.1 *Stieltjes transform and its properties*

If μ is a probability measure on $\mathbb{R}$, its **Stieltjes transform** is the function:

$$W_\mu(z) = \int_{\mathbb{R}} \frac{\mathrm{d}\mu(x)}{z-x}\,. \tag{11}$$

It is a holomorphic function[6] of $z \in \mathbb{C} \setminus \operatorname{supp} \mu$. It is an important tool because of the Stieltjes continuity theorem – see for instance [31]. In its most basic form:

5.1 THEOREM. *Let $(\mu_n)_n$ be a sequence of probability measures on $\mathbb{R}$, and μ another probability measure. μ_n converges to μ for the vague topology if and only if for all $z \in \mathbb{C} \setminus \mathbb{R}$, $W_{\mu_n}(z)$ converges to $W_\mu(z)$.*

The same theorem holds if $(\mu_n)_n$ is a sequence of random measures, by adding on both sides of the equivalence the mode of convergence "almost sure", "in probability", etc. Thus, the problem of checking the convergence of probability measures can thus be replaced with the – usually easier – problem of checking pointwise convergence of holomorphic functions. Let us give a few useful properties to handle the Stieltjes transform.

• Firstly, if μ is a measure which has moments up to order K, we have the asymptotic expansion:

$$W_\mu(z) = \frac{1}{z} + \sum_{k=1}^{K} \frac{m_k}{z^{k+1}} + o(z^{-(K+1)}), \qquad m_k = \int_{\mathbb{R}} x^k \,\mathrm{d}\mu(x)$$

valid when $|z| \to \infty$ and z remains bounded away from the support (if the support is $\mathbb{R}$, that means $|\operatorname{Im} z| \geq \delta$ for some fixed $\delta > 0$). So, the moments can be read off the expansion of $W_\mu(z)$ at infinity.

• Secondly, the Stieltjes transform can be given a probabilistic interpretation. We observe that, for $y \in \mathbb{R}$ and $\eta > 0$,

$$-\frac{1}{\pi} \operatorname{Im} W_\mu(y + \mathrm{i}\eta) = \int_{\mathbb{R}} \frac{\eta}{\pi} \frac{\mathrm{d}\mu(x)}{(y-x)^2 + \eta^2}$$

is the density – expressed in the variable y – of the convolution $\mu \star C_\eta$ of the initial measure μ with the Cauchy measure of width η:

$$C_\eta = \frac{\eta\,\mathrm{d}x}{\pi(x^2+\eta^2)}\,.$$

• Thirdly, the measure μ can be retrieved from its Stieltjes transform. Indeed, if f is a continuous function bounded by a constant $M > 0$, we know that:

$$\lim_{\eta \to 0} \int_{\mathbb{R}} \frac{\eta}{\pi} \frac{f(y)\,\mathrm{d}y}{(x-y)^2+\eta^2} = f(x)\,,$$

[6] The support $\operatorname{supp} \mu$ is the set of all points $x \in \mathbb{R}$ such that, for any open neighborhood U_x of x, $\mu[U_x] > 0$.

and actually the quantity inside the limit is bounded by M. So, by dominated convergence, we have:

$$\lim_{\eta\to 0^+} \int_{\mathbb{R}} f(y)\,\mathrm{d}(\mu \star C_b)(y) = \lim_{\eta\to 0^+} \int_{\mathbb{R}} \mathrm{d}\mu(x) \Big(\int_{\mathbb{R}} \frac{\eta}{\pi} \frac{f(y)\,\mathrm{d}y}{(x-y)^2+\eta^2} \Big) = \int_{\mathbb{R}} f(x)\,\mathrm{d}\mu(x)\,. \tag{12}$$

This means that, if μ has a density[7], this density is computed as the discontinuity on the real axis of the Stieltjes transform:

$$\mu(x) = \frac{W_\mu(x-\mathrm{i}0) - W_\mu(x+\mathrm{i}0)}{2\mathrm{i}\pi}\,\mathrm{d}x\,. \tag{13}$$

Note that there is a unique function $W(z)$ which is holomorphic in $\mathbb{C}\setminus\mathbb{R}$, has a given discontinuity on $\mathbb{R}$, and behaves likes $1/z$ when $|z|\to\infty$. Indeed, if $\tilde{W}$ was another such function, then $\tilde{W}-W$ would have no discontinuity on $\mathbb{R}$, hence would be holomorphic in $\mathbb{C}$. The growth condition implies that it decays at infinity, and by Liouville theorem, this implies that $\tilde{W}-W=0$.

Let us see how it works on a few examples.

• The Stieltjes transform of a Dirac mass located at x_0 is:

$$W(z) = \frac{1}{z-x_0}\,.$$

More generally, a simple pole at $z=x_0\in\mathbb{R}$ with residue r in $W_\mu(z)$ indicated that μ has a contribution from a Dirac mass r located at x_0.

• For the semi-circle law (10), we could use the definition (11) and compute the integral with the change of variable $x=\sigma(\zeta+1/\zeta)$ and complex analysis tricks. But there is a better way, relying on (13). Indeed, we are looking for a holomorphic function behaving like $1/z$ when $|z|\to\infty$, which has a discontinuity on $[-2\sigma,2\sigma]$ such that:

$$\forall x\in[-2\sigma,2\sigma],\qquad W(x+\mathrm{i}0)-W(x-\mathrm{i}0) = -\frac{\sqrt{x^2-4\sigma^2}}{\sigma^2}\,.$$

But we know that the squareroot takes a minus sign when one crosses the locus $[-2\sigma,2\sigma]$ where the quantity inside is negative, so its discontinuity is twice the squareroot. Therefore, the function $-\frac{1}{2\sigma^2}\sqrt{z^2-4\sigma^2}$ has the discontinuity we look for. It cannot be the final answer for $W(z)$, because of the condition $W(z)\sim 1/z$ when $|z|\to\infty$. But this can be achieved by adding a polynomial: it does not affect the holomorphicity and discontinuity, but can compensate the growth of the squareroot at infinity. One can check that:

$$W_{\mathrm{sc}}(z) = \frac{z-\sqrt{z^2-4\sigma^2}}{2\sigma^2} \tag{14}$$

[7] If μ has no density, (13) has to be interpreted in the weak sense (12).

has all the required properties, provided we choose the determination of the squareroot such that $\sqrt{z^2-4\sigma^2} \sim z$ when $|z| \to \infty$. By uniqueness, (14) must be the Stieltjes transform of μ_{sc}.

• Inspired by these two examples, the reader can show that the Stieltjes transform of the Marčenko-Pastur law is:

$$W_{MP}(z) = \frac{(1-\gamma)\sigma^2 + \gamma z - \gamma\sqrt{(z-a_+(\gamma))(z-a_-(\gamma))}}{2\sigma^2 z},$$

where the determination of the squareroot is fixed by requiring that:

$$\sqrt{(z-a_+(\gamma))(z-a_-(\gamma))} \sim z$$

when $|z| \to \infty$.

5.2 *$\mathcal{R}$-transform*

A closely related tool is the $\mathcal{R}$-transform. To simplify, we consider only measures μ for which the moments $m_k = \mu[x^k]$ exist for all $k \geq 0$. Let us consider the formal Laurent series:

$$\mathcal{W}_\mu(z) = \frac{1}{z} + \sum_{k\geq 1} \frac{m_k}{z^{k+1}}. \tag{15}$$

We shall use curly letters to distinguish the formal series from the holomorphic function $W_\mu(z)$. There exist a unique formal series:

$$\mathcal{R}_\mu(w) = \frac{1}{w} + \sum_{\ell \geq 1} \kappa_\ell \, w^{\ell-1} \tag{16}$$

such that:

$$\mathcal{R}_\mu(\mathcal{W}_\mu(z)) = z. \tag{17}$$

In other words, $\mathcal{R}_\mu$ is the functional inverse – at the level of formal series – of $\mathcal{W}_\mu$. So, we also have equivalently $\mathcal{W}_\mu(\mathcal{R}_\mu(w)) = 0$. If we declare that m_k has degree k, the κ_ℓ are homogeneous polynomials of degree ℓ in the $(m_k)_{k\geq 1}$. One can compute them recursively by replacing (15)-(16) in (17):

$$\begin{aligned}
\kappa_1 &= m_1, \\
\kappa_2 &= m_2 - m_1^2, \\
\kappa_3 &= m_3 - 3m_1m_2 + 2m_1^3, \\
\kappa_4 &= m_4 - 4m_1m_3 - 2m_2^2 + 10m_2m_1^2 - 5m_1^4, \ldots
\end{aligned}$$

The κ_ℓ are called **free cumulants**. They should not be confused with the better known **cumulants** $(c_\ell)_{\ell\geq 1}$, defined by:

$$\ln\left(1 + \sum_{k\geq 1} \frac{m_k \, t^k}{k!}\right) = \sum_{\ell \geq 1} \frac{c_\ell \, t^\ell}{\ell!}, \qquad t \to 0$$

We see on the first few values:

$$\begin{aligned} c_1 &= m_1\,, \\ c_2 &= m_2 - m_1^2\,, \\ c_3 &= m_3 - 3m_1m_2 + 2m_1^3\,, \\ c_4 &= m_4 - 4m_1m_3 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4\,, \ldots \end{aligned}$$

that $c_2 = \kappa_2$ and $c_3 = \kappa_3$, but this is accidental and in general the cumulants and free cumulants differ for $\ell \geq 4$.

5.3 *Asymptotic freeness*

In general, if A and B are two hermitian matrices, the knowledge of the spectrum of A and B is not enough to determine the spectrum of $A + B$ or $A \cdot B$. Indeed, when A and B do not commute, they cannot be diagonalized in the same basis.

It turns out that for large random matrices "in general position", knowing the spectrum of A and B is enough to reconstruct the spectrum of $A + B$, and the answer is elegantly expressed in terms of the $\mathcal{R}$-transform ; the theory is mainly due to Voiculescu around 1991 [35], in the more general context of C^* algebras. Explaining why this is true would bring us too far, but we aim at presenting the recipe, and illustrating some of its consequences.

We start by introducing several notions, first in a non-random context.

5.2 definition. If $(M_n)_n$ is a sequence of hermitian matrices of size n, we say that it has a limit distribution if there exists a probability measure μ with compact support such that $L^{(M_n)}$ converges to μ for the vague topology.

5.3 definition. Let $(A_n)_n$ and $(B_n)_n$ two sequences of hermitian matrices of size n, admitting as limit distributions respectively μ_A and μ_B. We say that $(A_n)_n$ and $(B_n)_n$ are **asymptotically free** if for any positive integers $r, m_1, m_1', \ldots, m_r, m_r'$, we have:

$$\lim_{n\to\infty} n^{-1} \operatorname{Tr}\Big\{ \prod_{i=1}^{r} (A_n^{m_i} - \mu_A[x^{m_i}] \cdot I_n)(B_n^{m_i'} - \mu_B[x^{m_i'}] \cdot I_n) \Big\} = 0\,, \tag{18}$$

where I_n is the identity matrix of size n, and the factors in the product are written from the left to the right with increasing i.

If we expand (18) and use it recursively, it implies that for asymptotically free matrices, the large n limit of the trace of arbitrary products of A_n and B_n can be computed solely in terms of the moments of μ_A and μ_B. In particular, the large n limit of $n^{-1}\operatorname{Tr}(A_n + B_n)^m$ or $n^{-1}\operatorname{Tr}(A_n \cdot B_n)^m$ can be computed solely in terms of μ_A and μ_B. Since measures with compact support are determined by their moments, we therefore understand that μ_A and μ_B should determine μ_{A+B} and $\mu_{A\cdot B}$. Finding the explicit formulas requires some combinatorial work. Focusing on the spectrum of the sum, the result is:

5.4 theorem. *If $(A_n)_n$ and $(B_n)_n$ are asymptotically free and have limit distributions μ_A and μ_B, then $(A_n + B_n)_n$ has a limit distribution μ_{A+B}, characterized by:*

$$\mathcal{R}_{\mu_{A+B}}(w) = \mathcal{R}_{\mu_A}(w) + \mathcal{R}_{\mu_B}(w) - \frac{1}{w}\,. \tag{19}$$

The last term $-\frac{1}{w}$ is there to ensure that the right-side is of the form $1/w + O(1)$ when $w \to 0$.

The relevance of this result in random matrix theory is illustrated by the following theorem of Voiculescu:

5.5 theorem. *Let $(A_n)_n$ and $(B_n)_n$ be two sequences of hermitian random matrices of size n. Assume that, for any n, A_n is independent of B_n, and for any unitary matrix Ω_n, $\Omega_n^{-1} A_n \Omega_n$ is distributed like A_n. Then, $(A_n)_n$ and $(B_n)_n$ are almost surely asymptotically free.*

In particular, if $L^{(A_n)}$ (resp. $L^{(B_n)}$) converges almost surely to a deterministic μ_A (resp μ_B) for the vague topology, using Stieltjes continuity theorem, one deduces that $L^{(A_n+B_n)}$ converges almost surely to a deterministic μ_{A+B} characterized by (19). To compute it, one has to compute the Stieltjes transforms $\mathcal{W}_{\mu_A}$ and $\mathcal{W}_{\mu_B}$, then compute their functional inverses $\mathcal{R}_{\mu_A}$ and $\mathcal{R}_{\mu_B}$, use (19), compute again the functional inverse $\mathcal{W}_{\mu_{A+B}}$, and finally reconstruct the measure μ_{A+B} from (13).

5.4 *The semi-circle law as a non-commutative CLT*

From Voiculescu's result, one can understand that the semi-circle law is an analog, in the non-commutative world, of the gaussian distribution arising when summing independent, identically distributed (i.i.d) real-valued random variables.

Let $(A_n^{(j)})_{1 \le j \le N}$ be i.i.d, centered random matrices, whose distribution is invariant under conjugation by a unitary matrix. We assume that the empirical measure of $A_n^{(1)}$ converges almost surely to μ_A for the vague topology. It follows from a slight generalization of Voiculescu's theorem that the family $((A_n^{(j)})_{1 \le j \le N})_n$ is asymptotically free – this is defined like in Definition 5.3, except that one uses arbitrary sequences of letters $A^{(j_1)} \cdots A^{(j_s)}$ instead of arbitrary sequences of letters $ABABAB \cdots$. Let us consider:

$$S_n^{(N)} = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} A_n^{(j)}\,.$$

Theorem 5.4 has an obvious generalization to this case: for any $N \ge 1$, $(S_n^{(N)})_n$ has a limit distribution $\mu_{S^{(N)}}$ when $n \to \infty$, which is characterized by:

$$\mathcal{R}_{\mu_{S^{(N)}}} = N \mathcal{R}_{\mu_{A/\sqrt{N}}}(w) - \frac{N-1}{w}\,.$$

Playing with the functional equation (17), one easily find what is the effect of a rescaling on the $\mathcal{R}$-transform:

$$\mathcal{R}_{\mu_A/\sqrt{N}}(w) = N^{-1/2}\mathcal{R}_{\mu_A}(N^{-1/2}w)\,.$$

Since $A_n^{(1)}$ is centered, the first moment of μ_A vanishes. Denoting σ_A^2 the variance of μ_A, we can write:

$$\mathcal{R}_{\mu_A}(w) = \frac{1}{w} + \sigma^2\, w + \sum_{\ell\geq 2} \kappa_{\ell+1}\, w^\ell\,,$$

and therefore:

$$\mathcal{R}_{\mu_{S^{(N)}}}(w) = \frac{1}{w} + \sigma^2 w + \sum_{\ell\geq 2} N^{(1-\ell)/2}\,\kappa_{\ell+1} w^\ell \underset{N\to\infty}{\longrightarrow} \frac{1}{w} + \sigma^2 w \tag{20}$$

The functional inverse of $\mathcal{R}_\infty(w) = \frac{1}{w} + \sigma^2\, w$ can be readily computed as it is solution of a quadratic equation:

$$\mathcal{R}_\infty(\mathcal{W}_\infty(z)) = z \qquad \Longleftrightarrow \qquad \mathcal{W}_\infty(z) = \frac{z - \sqrt{z^2 - 4\sigma^2}}{2\sigma^2 z}\,.$$

Note that the determination of the squareroot is fixed by requiring that the formal series $\mathcal{W}_\infty(z)$ starts with $1/z + O(1/z)$. We recognize the Stieltjes transform (14) of the semi-circle law μ_{sc} with variance σ^2. Using Stieltjes continuity theorem, one can deduce that $\mu_{S^{(N)}}$ converges for the vague topology to μ_{sc} when $N \to \infty$. It is remarkable that the limit distribution for $S_n^{(N)}$ when $n, N \to \infty$ does not depend on the details of the summands $A_n^{(j)}$.

Actually, the mechanism of the proof is similar to that of the central limit theorem, provided one replaces the notion of Fourier transform (which is multiplicative for sum of independent real-valued random variables) with the notion of $\mathcal{R}$-transform (which is additive for the sum asymptotically free random matrices). In both cases, the universality of the result – as well as the occurrence of the gaussian distribution/the semi-circle law – comes from the fact that, when the number of summands N goes to infinity, only the second moment survives in the formula characterizing the distribution.

5.5 *Perturbation by a finite rank matrix*

We show[8] how simple computations with the $\mathcal{R}$-transform give insight into the effect of a finite rank perturbation on the spectrum of a GUE matrix. This gives a good qualitative idea of the effect of perturbations on more general random matrices. We will state in Section 6 a complete theorem for Wishart matrices.

[8] The example we present is inspired by Bouchaud.

So, let A_n be a GUE matrix of size n with variance σ^2, and consider:

$$S_n = A_n + B_n, \qquad B_n = \mathrm{diag}(\underbrace{\Lambda, \dots, \Lambda}_{m \text{ times}}, \underbrace{0, \dots, 0}_{n-m \text{ times}})$$

for $\Lambda > 0$. We set:

$$\epsilon = \frac{m}{n}$$

and would like the study the limit $n \to \infty$, and then ϵ is small. As we have seen, the distribution of A_n is invariant under conjugation by a unitary matrix, and it has the semi-circle law as limit distribution. B_n is deterministic, therefore independent of A_n, and it admits a limit distribution given by:

$$\mu_B = (1-\epsilon)\delta_0 + \epsilon\delta_\Lambda \,. \tag{21}$$

This falls in framework of Voiculescu's theorem, so S_n has a limit distribution μ_S. To compute it, we first write down the Stieltjes transform:

$$W_{\mu_B}(z) = \frac{1-\epsilon}{z} + \frac{\epsilon}{z-\Lambda}\,,$$

and solving for the functional inverse:

$$\mathcal{R}_{\mu_B}(w) = \frac{1}{2}\left[\frac{1}{w} + \Lambda + \sqrt{\Big(\frac{1}{w} - \Lambda)^2 + \frac{4\epsilon\Lambda}{w}}\right].$$

Therefore, we add to it the $\mathcal{R}$-transform (20) of the semi-circle law minus $1/w$, and we can expand when $\epsilon \to 0$:

$$\begin{aligned} \mathcal{R}_{\mu_S}(w) &= \sigma^2 w + \frac{1}{2}\left[\frac{1}{w} + \Lambda + \sqrt{\Big(\frac{1}{w} - \Lambda)^2 + \frac{4\epsilon\Lambda}{w}}\right] \\ &= \frac{1}{w} + \sigma^2 w + \frac{\epsilon\Lambda}{1-\Lambda w} + O(\epsilon^2)\,. \end{aligned} \tag{22}$$

The Stieltjes transform of μ_S will satisfy:

$$z = \frac{1}{W_{\mu_S}(z)} + \sigma^2 W_{\mu_S}(z) + \frac{\epsilon\Lambda}{1-\Lambda W_{\mu_S}(z)} + O(\epsilon^2)\,. \tag{23}$$

At leading order in ϵ, μ_S the semi-circle law. Let us have a look at the first subleading correction. Qualitatively, two situations can occur.

• If $W_{\mathrm{sc}}(z) = 1/\Lambda$ admits a solution $z = z_\Lambda$ on the real axis outside of the support $K_\sigma = [-2\sigma, 2\sigma]$ of μ_S, the $O(\epsilon)$ correction to W_{μ_S} has a singularity outside K_σ, which is the sign that μ_S has some mass outside K_σ. If such a real-valued z_Λ exists, we must have:

$$\frac{1}{\Lambda} = \frac{z_\Lambda - \sqrt{z_\Lambda^2 - 4\sigma^2}}{2\sigma^2} \leq \frac{z_\Lambda - (z_\Lambda - 2\sigma)}{2\sigma^2} \leq \frac{1}{\sigma}\,.$$

Conversely, if the condition $\Lambda > \sigma$ is met, then there exists a unique such z_Λ,

given by:

$$z_\Lambda = \Lambda + \frac{\sigma^2}{\Lambda}\,.$$

One can then show solving (23) perturbatively that $W_{\mu_S}(z)$ has a simple pole at $z = z_\Lambda + o(1)$, with residue $\epsilon + o(\epsilon)$. This means that μ_S has a Dirac mass ϵ at z_Λ. In other words, if Λ is above the threshold σ, a fraction ϵ of eigenvalues – i.e. $m = \mathrm{rank}(B_n)$ eigenvalues – detach from the support. Even for ϵ arbitrarily small but non-zero, the maximum eigenvalue is now located at $z_\Lambda > 2\sigma$ instead of 2σ for a GUE matrix.
• If $\Lambda \leq \sigma$, the singularities of $W_{\mu_S}(z)$ remain on K_σ, and therefore the density of μ_S is a small perturbation of the semi-circle, not affecting the position of the maximum eigenvalue.

One should note that the value of the threshold $\Lambda_* = \sigma$ is located in the bulk of the support. We will justify in Section 7.1 the loose statement that:

"eigenvalues of random matrices repel each other"

This allows an interpretation of the above phenomenon. If we try to add to a random matrix a deterministic matrix with m eigenvalues Λ, they will undergo repulsion of the eigenvalues that were distributed according to the distribution of A (here, the semi-circle). If the m Λ's feel too many eigenvalues of A to their left – here it happens precisely when $\Lambda > \sigma$ – they will be kicked out from the support, to a location z_Λ further to the right of the support. If $\Lambda < \sigma$, the Λ's feel the repulsion of enough eigenvalues to their right and to their left to allow for a balance, and thus we just see a small deformation of the semi-circle law, keeping the same support in first approximation.

6 Wishart matrices with perturbed covariance

Lecture 3 (1h30)
August 7th, 2015

The same phenomenon was analyzed for complex Wishart matrices by Baik, Ben Arous and Péché [2], and is now called the **BBP phase transition**. The result also holds for real Wishart matrices [3]. We consider a Wishart matrix M of size p, with n degrees of freedom, and covariance $K = \mathrm{diag}(\Lambda^2, \sigma^2, \dots, \sigma^2)$. This is a perturbation of the null model with covariance $\mathrm{diag}(\sigma^2, \dots, \sigma^2)$.

6.1 theorem. *Assume $n, p \to \infty$ while n/p converges to γ, and define:*

$$\Lambda_* = \sigma(1 + \gamma^{-1/2}) .$$

- *If $\Lambda \in (0, \Lambda_*)$, Theorem 3.3 continues to hold: $\lambda_1^{(M)}$ converges almost surely to $a_+(\gamma)$, and the fluctuations at scale $p^{-2/3}$ follow the Tracy-Widom law.*
- *If $\Lambda \in (\Lambda_*, +\infty)$, we have almost sure convergence of the maximum:*

$$\lambda_1^{(M)} \longrightarrow z_\Lambda := \sigma\Lambda\Big(1 + \frac{\sigma}{\gamma(\Lambda - \sigma)}\Big) ,$$

  and the random variable

$$\frac{p^{1/2}}{\sigma\Lambda}\Big(\frac{1}{\gamma} - \frac{\sigma^2}{\gamma^2(\Lambda - \sigma)^2}\Big)^{1/2}\big\{\lambda_1^{(M)} - z_\Lambda\big\}$$

  describing fluctuations at scale $p^{-1/2}$, converges in law to a Gaussian with variance 1.

When Λ approaches Λ_* at a rate depending on p, the maximum eigenvalues converges to $a_+(\gamma)$, but its fluctuations follows a new distribution, that interpolates between Tracy-Widom and Gaussian laws.

For application in statistics, Λ can be thought as a trend in empirical data. One may wonder if the trend can be identified from a PCA analysis. The theorem shows that the answer is positive only if the trend is strong enough – i.e. $\Lambda > \Lambda_*$. As for perturbation of the GUE, the threshold Λ_*^2 lies inside the support of the Marčenko-Pastur law.

Although more interesting for statistics, the case of real Wishart matrices was only tackled in 2011 by Bloemendal and Virág[9], with similar conclusions. The reason is that, in the complex case, we will see in Section 8.3 that algebraic miracles greatly facilitates the computations, which boil down to analyzing the asymptotic behavior of a sequence of orthogonal polynomials. This can be done with the so-called Riemann-Hilbert steepest descent analysis, developed by Deift, Zhou and coauthors in the 90s – for an introduction, see [9] – and this is the route taken by BBP.

[9] Actually, their method relate the distributions for the fluctuations of the maximum of perturbed GOE or GUE to the probability of explosion of the solution of second order stochastic differential equation. In the unperturbed case, they also obtained characterizations of the same nature for the Tracy-Widom laws. This is a beautiful result fitting in the topic of the summer school, however at a more advanced level compared to the background provided at the school.

7 From matrix entries to eigenvalues

7.1 *Lebesgue measure and diagonalization*

We would like to compute the joint distribution of eigenvalues of a symmetric or hermitian random matrix. For this purpose, we basically need to perform a change of variables in integrals of the form $\int \mathrm{d}M\, f(M)$, hence to compute the determinant of the Jacobian of this change of variable. Although some details have to be taken care of before arriving to that point, the core of the computation is easy and concentrated in (27) and the evaluation of the determinant.

First consider the case of symmetric matrices. Let $\mathcal{O}_n$ be the set of orthogonal $n \times n$ matrices, i.e. satisfying $\Omega^T \Omega = I_n$. Since any symmetric matrix can be diagonalized by an orthogonal matrix, the $\mathcal{C}^\infty$ map:

$$(24)\quad M : \begin{array}{ccc} \mathcal{O}_n \times \mathbb{R}^n & \longrightarrow & \mathcal{S}_n \\ (\Omega, \lambda_1, \dots, \lambda_n) & \longmapsto & \Omega\, \mathrm{diag}(\lambda_1, \dots, \lambda_n)\Omega^{-1} \end{array}$$

is surjective. However, the map is not injective, so we cannot take (24) as an admissible change of variable. Indeed, if:

$$M = \Omega \mathrm{diag}(\lambda_1, \dots, \lambda_n)\Omega^{-1} = \tilde{\Omega} \mathrm{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_n)\tilde{\Omega}^{-1},$$

then there exists a permutation $\sigma \in \mathfrak{S}_n$ and an orthogonal matrix D that leaves stable the eigenspaces of M such that:

$$(25)\quad \tilde{\Omega} = \Omega D, \qquad \tilde{\lambda}_i = \lambda_{\sigma(i)}\,.$$

To solve this issue, we first restrict to the subset $(\mathcal{S}_n)_\Delta$ consisting of symmetric matrices with pairwise distinct eigenvalues. This is harmless since $(\mathcal{S}_n)_\Delta$ is an open dense subset of $\mathcal{S}_n$, hence its complement has Lebesgue measure 0. Then, two decompositions are related by (25) with D being a diagonal orthogonal matrix, and this forces the diagonal entries to be ± 1. So, let us mod out the left-hand side of (24) by $\{\pm 1\}^n$. Then, we can kill the freedom of permuting the λ_i's by requiring that λ_i decreases with i. Denoting:

$$(\mathbb{R}_n)_\Delta = \left\{(\lambda_1, \dots, \lambda_n) \in \mathbb{R}^n, \qquad \lambda_1 > \lambda_2 > \dots > \lambda_n\right\},$$

we finally obtain an invertible map:

$$(26)\quad M : \begin{array}{ccc} (\mathcal{O}_n/\{\pm 1\}^n) \times (\mathbb{R}^n)_\Delta & \longrightarrow & (\mathcal{S}_n)_\Delta \\ (\Omega, \lambda_1, \dots, \lambda_n) & \longmapsto & \Omega\, \mathrm{diag}(\lambda_1, \dots, \lambda_n)\Omega^{-1} \end{array}$$

and one can show that it is a $\mathcal{C}^\infty$ diffeomorphism – i.e. an admissible change of variable.

To be more explicit, we have to choose coordinates on $\mathcal{O}_n$. In the vicinity of $I_n \in \mathcal{O}_n$, we can choose as coordinates the entries $(\omega_{ij})_{1 \le i < j \le n}$ of an antisymmetric matrix ω, which parametrizes an orthogonal matrix by the formula $\Omega = \exp(\omega)$. And in $\mathcal{S}_n$, we remind that we had chosen as coordinates the

entries $(M_{ij})_{1\leq i\leq j\leq n}$. Then, we know that:

$$\mathrm{d}M = 2^{-n} \prod_{1\leq i<j\leq n} \mathrm{d}\omega_{ij} \prod_{i=1}^{n} \mathrm{d}\lambda_i \, \mathcal{J}(\lambda,\omega)\,,$$

where the 2^{-n} comes from the quotient by $\{\pm 1\}^n$, and it remains to compute the Jacobian determinant:

$$\mathcal{J}(\lambda,\omega) = \left| \det \left[\begin{array}{cc} \frac{\partial M_{ij}}{\partial \omega_{kl}} & \frac{\partial M_{ij}}{\partial \lambda_k} \\ {\scriptstyle 1\leq k<l\leq n} & {\scriptstyle 1\leq k\leq n} \end{array} \right]_{1\leq i\leq j\leq n} \right|$$

First, we remind that the Lebesgue measure is invariant under conjugation of M by an orthogonal matrix. We can thus evaluate the derivatives at $\omega = 0$ (i.e. $\Omega = I_n$) and find:

$$\mathrm{d}M_{ij} = [\mathrm{d}\omega, \Lambda]_{ij} + \mathrm{d}\Lambda_i \delta_{ij} = \mathrm{d}\omega_{ij}(\lambda_i - \lambda_j) + \mathrm{d}\lambda_i \delta_{ij} \tag{27}$$

Therefore, the matrix in the Jacobian is diagonal: in the first block $1 \leq i < j \leq n$ and $1 \leq k < l \leq n$, the diagonal elements $(i,j) = (k,l)$ are $(\lambda_i - \lambda_j)$, and in the second block, the diagonal elements are just 1. Therefore:

$$\mathcal{J}(\lambda,0) = \prod_{1\leq i<j\leq n} |\lambda_j - \lambda_i|$$

We can repeat all steps for hermitian matrices. $\mathcal{O}_n$ should be replaced with the set $\mathcal{U}_n$ of unitary matrices, i.e. satisfying $(\Omega^T)^*\Omega = I_n$. The map (24) now sends $\mathcal{U}_n \times \mathbb{R}^n$ to $\mathcal{H}_n$. It is not surjective, but if we restrict to the set $(\mathcal{H}_n)_\Delta$ of hermitian matrices with pairwise distinct eigenvalues, the only freedom is to have (25) with D a diagonal matrix whose entries are complex numbers of unit norm ; we denote $\mathcal{U}_1^n$ the group of such matrices. Then, we obtain an admissible change of variable:

$$\left(\mathcal{U}_n/\mathcal{U}_1^n\right) \times (\mathbb{R}^n)_\Delta \simeq (\mathcal{H}_n)_\Delta\,. \tag{28}$$

As coordinates on $\mathcal{U}_n$ near I_n, we can take the real and imaginary parts of the entries $(\omega_{ij})_{1\leq i<j\leq n}$ of a matrix ω such that[10] $(\omega^T)^* = -\omega$, parametrizing a unitary matrix by the formula $\Omega = \exp(\omega)$. The formula (27) for the differential does not change but we have now twice many coordinates: the Jacobian matrix is still diagonal, and the diagonal entries corresponding to derivative with respect to $\mathrm{Re}\,\omega_{ij}$ and to $\mathrm{Im}\,\omega_{ij}$ both evaluate to $(\lambda_i - \lambda_j)$. Thus, the Jacobian determinant reads:

$$\mathcal{J}(\lambda,0) = \prod_{1\leq i<j\leq n} |\lambda_j - \lambda_i|^2\,.$$

[10]Such a matrix is called "antihermitian".

There is a last step about which we will be brief: this result – valid at $\Omega = I_n$ – has to be transported to any point of $\mathcal{S}_n$ (or $\mathcal{H}_n$) by conjugating with an $\mathcal{O}_n$ (resp $\mathcal{U}_n$) matrix. Of course, this does not affect the eigenvalue dependence of the Jacobian factor. The result makes appear the Haar measure on $\mathcal{O}_n$ (resp. $\mathcal{U}_n$): this is the unique probability measure which is invariant under left and right multiplication by an orthogonal (resp. unitary) matrix. We denote $\mathrm{d}\nu(\Omega)$ the measure induced by the Haar measure on the quotient $\mathcal{O}_n/\{\pm 1\}^n$ (resp. $\mathcal{U}_n/\mathcal{U}_1^n$).

7.1 theorem. *Under the change of variable* (26) *or* (28)*, we have:*

$$\mathrm{d}M = c_{\beta,n}\,\mathrm{d}\nu(\Omega)\prod_{i=1}^{n}\mathrm{d}\lambda_i \prod_{1\le i<j\le n}|\lambda_j-\lambda_i|^\beta$$

for some (explicitly computable) constant $c_{\beta,n} > 0$*.*

7.2 *Repulsion of eigenvalues*

As a consequence, if M is a random symmetric (resp. hermitian) matrix whose p.d.f. of entries is $\mathrm{d}M\,F(M)$, and f is invariant under conjugation by an orthogonal (resp. unitary) matrix, then $F(M)$ is actually a function $f(\lambda_1,\dots,\lambda_n)$ of the eigenvalues only, and the joint p.d.f of the eigenvalues of M is proportional to:

$$Z_{n,\beta}^{-1}\prod_{1\le i<j\le n}\big|\Delta(\lambda_1,\dots,\lambda_n)\big|^\beta f(\lambda_1,\dots,\lambda_n)\,, \tag{29}$$

with:

$$\Delta(\lambda_1,\dots,\lambda_n) = \prod_{1\le i<j\le n}(\lambda_j-\lambda_i)\,, \tag{30}$$

and the constant $Z_{n,\beta}$ is such that the integral of (29) against the Lebesgue measure over $\mathbb{R}^n$ evaluates to 1. Because of the factor $|\Delta(\lambda_1,\dots,\lambda_n)|^\beta$ the probability that two eigenvalues are close to each other is small: the eigenvalues of a random matrix usually repel each other. The intensity of the repulsion is measured by the parameter β, which is fixed by the type of the matrix (symmetric or hermitian).

7.2 lemma. (30) *is the* ***Vandermonde determinant****:*

$$\Delta(\lambda_1,\dots,\lambda_n) = \det\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix}.$$

Proof. Let us denote $D(\lambda_1,\dots,\lambda_n)$ the determinant in the right-hand side. It is a polynomial function of λ_i, of degree at most $n-1$, which admits the $n-1$ roots $\lambda_i = \lambda_j$ indexed by $j \neq i$. Therefore, we can factor out successively all

the monomials that occur in Δ, and find:

$$D(\lambda_1, \dots, \lambda_n) = c_n \, \Delta(\lambda_1, \dots, \lambda_n) \tag{31}$$

for some constant c_n. We prove by induction that $c_n = 1$. This is obviously true for $n = 1$. If this is true for $(n-1)$, we expand the determinant of size n with respect to its last column, and find that the coefficient of λ_n^{n-1} is $D(\lambda_1, \dots, \lambda_{n-1})$. Comparing with (31) and the induction hypothesis, we deduce that $c_n = 1$. □

7.3 lemma. *For any sequence $(Q_m)_{m \geq 0}$ of polynomials of degree m with leading coefficient 1:*

$$\Delta(\lambda_1, \dots, \lambda_n) = \det_{1 \leq i,j \leq n} \left[Q_{i-1}(\lambda_j) \right].$$

Proof. By adding linear combinations of the $(n-1)$ first lines to the last line, one can actually replace λ_j^{n-1} in the last line by $Q_{n-1}(\lambda_j)$ for any polynomial Q_{n-1} of degree $n-1$ with leading coefficient 1. Repeating this procedure successively for the lines $(n-1)$, $(n-2)$, etc. establishes the claim. □

7.3 *Eigenvalue distribution of Wishart matrices*

The result for Wishart matrices was obtained almost simultaneously in 1939 by [14, 17, 21, 28].

7.4 theorem. *If M is a real ($\beta = 1$) or complex ($\beta = 2$) Wishart matrix with covariance $K = \mathrm{diag}(\sigma^2, \dots, \sigma^2)$, of size p with n degrees of freedom, the joint p.d.f of its eigenvalues is:*

$$Z_{n,\beta}^{-1} \prod_{1 \leq i < j \leq n} |\lambda_i - \lambda_j|^\beta \prod_{i=1}^n \lambda_i^{\frac{\beta}{2}(n-p)+\frac{\beta-2}{2}} \exp\Big(- \frac{n\beta}{2\sigma^2} \lambda_i \Big) \tag{32}$$

for an (explicitly computable) normalization constant $Z_{n,\beta}^{-1}$.

Proof. The proof is a bit more involved than in Section 7.1, and was omitted during the lectures. It uses a change of variable in three steps, the last one being already given by Theorem 7.1. We give the details for the case of real Wishart matrices.

• First, we consider X as a matrix of p vectors in $\mathbb{C}^n$, which we can orthogonalize. This produces in a unique way a matrix Ω of size $n \times p$, such that:

$$\Omega^T \Omega = I_p \, . \tag{33}$$

and a lower-triangular matrix L of size $p \times p$ with positive diagonal entries, such that:

$$X = \Omega L \, . \tag{34}$$

The Lebesgue measure $\mathrm{d}X$ is invariant under multiplication to the left by an orthogonal matrix of size n, thus it is enough to evaluate the Jacobian at Ω equals:

$$\Omega^0 = \begin{bmatrix} I_{p,p} \\ 0_{n-p,p} \end{bmatrix},$$

where $0_{m,p}$ is the matrix of size $m \times p$ filled with 0's.

We need to fix local coordinates on the tangent space at Ω^0 of the set $\mathcal{O}_{n,p}$ of matrices Ω satisfying (33). For example, we can choose the entries Ω_{kl} with $1 \leq k < l \leq p$, and the Ω_{kl} with $k \geq p+1$ and $1 \leq l \leq p$. The remaining Ω_{kl} with $1 \leq l < k \leq p$ are then determined by (33), and infinitesimally around Ω^0 we find for these indices $\Omega_{kl} = -\Omega_{lk}$. The dimension of $\mathcal{O}_{n,p}$ is thus $p(p-1)/2 + p(n-p)$. For the matrix L, we naturally choose as coordinates its non-zero entries L_{kl} indexed by $1 \leq l \leq k \leq p$ – the space of L's has dimension $p(p+1)/2$. This is consistent with the dimension of the space of X's:

$$np = \frac{p(p+1)}{2} + \frac{p(p-1)}{2} + p(n-p)\,.$$

Now, we compute the differential of (34):

$$\mathrm{d}X_{ij} = \delta_{ik}\delta_{jl}\mathrm{d}L_{kl} + \mathrm{d}\omega_{kl}\delta_{ik}\delta_{k>l} - \mathrm{d}\omega_{kl}L_{kl}\delta_{il}\delta_{k<l} + \mathrm{d}\omega_{kl}L_{lj}\delta_{ik}\delta_{k>p}\,.$$

A careful look at the indices shows that the Jacobian matrix is of the form:

$$\begin{aligned}\mathcal{J}(L,\Omega) &= \left| \det \underset{1\leq l\leq k\leq p}{\underset{1\leq k<l\leq p}{\underset{k\geq p+1}{}}} \begin{bmatrix} \frac{\partial X_{ij}}{\partial L_{kl}} & \frac{\partial X_{ij}}{\partial \omega_{kl}} & \frac{\partial X_{ij}}{\partial \omega_{kl}} \end{bmatrix} \right| \\ &= \left| \det \underset{1\leq l\leq k\leq p}{\underset{1\leq k<l\leq p}{\underset{k\geq p+1}{}}} \begin{bmatrix} I & * & 0 \\ 0 & U & 0 \\ 0 & 0 & U' \end{bmatrix} \begin{matrix} 1\leq j\leq i\leq p \\ 1\leq i<j\leq p \\ i\geq p+1 \end{matrix} \right|\end{aligned}$$

with U and U' upper triangular matrices with respect to the lexicographic order on the double-indices (i,j). Besides, the diagonal elements of U and U' at position $(i,j) = (k,l)$ are L_{jj}. So, the determinant evaluates to:

$$\mathcal{J}(L,\Omega) = \prod_{j=1}^{p} L_{jj}^{n-p+j-1},$$

and we have:

$$\mathrm{d}X = \mathrm{d}\nu(\Omega) \prod_{1\leq j\leq i\leq p} \mathrm{d}L_{ij} \prod_{j=1}^{p} L_{jj}^{n-p+j-1}, \tag{35}$$

where $\mathrm{d}\nu(\Omega)$ is the measure on $\mathcal{O}_{n,p}$ obtained by transporting the volume element of the ω's from Ω^0 to any point in $\mathcal{O}_{n,p}$.

• Next, we change variables from L to M:

$$M = n^{-1}\, X^T X = n^{-1}\, L^T L\,.$$

The differential is:

$$\mathrm{d}M_{ij} = n^{-1}\Big(\delta_{lj} L_{ki} + \delta_{li} L_{kj}\Big)\,,$$

and we must compute the Jacobian:

$$\tilde{\mathcal{J}}(L) = \left| \det \left[\frac{\partial M_{ij}}{\partial L_{kl}} \right]_{1\le l\le k\le p}^{1\le j\le i\le p} \right|$$

If we put on couples (i,j) the lexicographic order, we observe that the Jacobian matrix is upper-triangular, with entries $n^{-1}(\delta_{jj}L_{ii} + \delta_{ij}L_{jj})$ on the diagonal with double index (i,j). Therefore:

$$\mathrm{d}M = \mathrm{d}L\, \tilde{\mathcal{J}}(L)\,, \qquad \tilde{\mathcal{J}}(L) = n^{-p(p+1)/2}\, 2^p \prod_{j=1}^{p} L_{jj}^{j}\,. \tag{36}$$

• Combining (35) and (36) yields:

$$\mathrm{d}X = c_{n,p}\, \mathrm{d}\nu(\Omega)\,, \mathrm{d}M \prod_{j=1}^{p} L_{jj}^{n-p-1}$$

and we rewrite:

$$\begin{aligned} \prod_{j=1}^{n} L_{jj}^{n-p-1} &= \det(L)^{n-p-1} = \det(L^T L)^{(n-p-1)/2} \\ &= n^{p(p+1)/2}\, \det(M)^{(n-p-1)/2} = n^{p(p+1)/2} \prod_{j=1}^{n} \lambda_j^{(n-p-1)/2}\,. \end{aligned}$$

Finally, we use Theorem 7.1 to obtain the announced result (32) in the case $\beta = 1$.

• The case of complex Wishart matrices is treated similarly, with $\mathcal{O}_{n,p}$ being replaced by the set $\mathcal{U}_{n,p}$ of $n \times p$ matrices Ω such that $(\Omega^T)^*\Omega = I_p$. □

8 Exact computations in invariant ensembles

8.1 *Invariant ensembles*

The Gaussian ensembles and the Wishart ensembles are special cases of the **invariant ensembles**. These are symmetric (resp. hermitian) random matrices M of size n, whose distribution of entries is of the form:

$$Z_{n,\beta}^{-1}\, \mathrm{d}M\, \exp\Big(-\frac{n\beta}{2}\operatorname{Tr} V(M)\Big)\,. \tag{37}$$

The function V is assumed to grow fast enough at infinity – e.g. V is a polynomial with positive leading coefficient – so that (37) has finite mass on $\mathcal{S}_n$ or $\mathcal{H}_n$, and we tune $Z_{n,\beta}^{-1}$ so that this mass is 1. Theorem 7.1 implies that the joint p.d.f of the eigenvalues[11] is:

$$Z_{n,\beta}^{-1} \prod_{1\le i<j\le n} |\lambda_i-\lambda_j|^\beta \prod_{i=1}^n \exp\Big\{-\frac{n\beta}{2}V(\lambda_i)\Big\}\,. \tag{38}$$

The Wishart ensembles – in which the size is denoted p instead of n – correspond to the case:

$$V(x) = -\frac{x}{\sigma^2} + \Big[\gamma - 1 + \frac{1}{p}\Big(1-\frac{2}{\beta}\Big)\Big]\ln x, \qquad \gamma = n/p\,, \tag{39}$$

and the Gaussian ensembles to:

$$V(x) = \frac{x^2}{2\sigma^2}\,.$$

Note that the distribution (38) makes sense for any value of $\beta > 0$. When β increases starting from 0, they provide an interpolating model from independent random variables to strongly correlated (repulsive) random variables, called the **β-ensembles**.

Equation 38 still contains too much information. We would like to answer questions like: what is the probability that one eigenvalue falls into a given interval ? In other words, we want to compute the marginals of the distribution (38). Surprisingly, for $\beta = 1$ and $\beta = 2$, this can be performed exactly, using tricks mainly discovered by Gaudin and Mehta in the early 60s. We will stick to the case $\beta = 2$, for which the computations are in fact much simpler. And since for the moment we will be occupied with exact computations, it is convenient to use a notation $W(\lambda_i)$ instead of $(n\beta/2)V(\lambda_i)$ in (38).

[11] Contrarily to the previous sections, in (38) the eigenvalues are not assumed to be ordered. When we need to consider the maximum eigenvalue, we shall use the notation $\lambda_{\max}$.

8.2 *Partition function*

Prior to any computation, it is useful to evaluate the normalization constant, also called **partition function**

$$Z_n = \int_{\mathbb{R}^n} \prod_{1\le i<j\le n} \big|\Delta(\lambda_1,\dots,\lambda_n)\big|^2 \prod_{i=1}^n e^{-W(\lambda_i)}\,.$$

This can be done in terms of the orthogonal polynomials $(P_n)_{n\ge 0}$ for the measure $\mathrm{d}x\, e^{-W(x)}$ on $\mathbb{R}$. More precisely, consider the scalar product on the space of real-valued polynomials:

$$\langle f, g\rangle = \int_{\mathbb{R}} f(x)\, g(x)\, e^{-W(x)}\,\mathrm{d}x\,. \tag{40}$$

The orthogonalization of the canonical basis $(x^n)_{n\ge 0}$ for the scalar product (40) determines a unique sequence $(p_n)_{n\ge 0}$ of polynomials with the following properties:

- P_n has degree n and starts with $x^n + \cdots$.
- For any $n, m \ge 0$, $\langle P_n, P_m\rangle = \delta_{nm} h_n$ for some constant $h_n > 0$.

8.1 theorem. $\boxed{Z_n = n! \prod_{m=0}^{n-1} h_m}$

Proof. Let $(Q_m)_{m\ge 0}$ be an arbitrary sequence of polynomials of degree m with leading coefficient 1, use the representation of Lemma 7.3 for Vandermonde determinant, and expand the determinants:

$$Z_n = \sum_{\sigma,\tau\in\mathfrak{S}_n} \mathrm{sgn}(\sigma)\mathrm{sgn}(\tau) \int_{\mathbb{R}^n} \prod_{i=1}^n Q_{\sigma(i)-1}(\lambda_i)\, Q_{\tau(i)-1}(\lambda_i)\, e^{-W(\lambda_i)}\mathrm{d}\lambda_i\,.$$

We observe that, in each term, the integral over $\mathbb{R}^n$ factors into n integrals over $\mathbb{R}$. Then, i is a dummy index for the product, and we can also rename it $\tau^{-1}(i)$. Since the signatures satisfies $\mathrm{sgn}(\sigma)\mathrm{sgn}(\tau) = \mathrm{sgn}(\sigma\tau^{-1})$, we shall change variables in the sum and set $\tilde\sigma = \sigma\tau^{-1}$. The summands only depend on $\tilde\sigma$, and it remains a sum over a permutation, which produces a factor of $n!$. So:

$$\begin{aligned} Z_n &= n! \sum_{\tilde\sigma\in\mathfrak{S}_n} \prod_{i=1}^n \Big[\int_{\mathbb{R}} Q_{\tilde\sigma(i)-1}(x)\, Q_{i-1}(x)\, e^{-W(x)}\mathrm{d}x\Big] \\ &= n! \det_{1\le i,j\le n} \Big[\int_{\mathbb{R}} Q_{i-1}(x) Q_{j-1}(x)\, e^{-W(x)}\Big]\,, \end{aligned} \tag{41}$$

where, in the last line, we have used the multilinearity of the determinant. Now, if we choose $(Q_m)_{m\ge 0}$ to be the orthogonal polynomials for the scalar product (40), the matrix in the determinant becomes diagonal. This entails the result. □

8.3 *Marginals of eigenvalue distributions*

Jánossy densities

If M is a random hermitian matrix, we define the k-point **Jánossy densities** $\rho_n^{(k)}(x_1, \dots, x_k)$, as the functions such that, for any pairwise disjoint measurable sets $A_1, \dots, A_k$:

$$\mathbb{P}\Big[\exists i_1, \dots, i_k, \qquad \lambda_{i_j} \in A_j\Big] = \int_{A_1 \times \cdots \times A_k} \rho_n^{(k)}(x_1, \dots, x_k) \prod_{i=1}^{k} \mathrm{d}x_i \,. \tag{42}$$

The $\rho_n^{(k)}$ can be considered as a probability density – in particular they are non-negative – except that their total integral is not 1. Since the eigenvalues are not ordered in (42), $\rho_n(k)$ is a symmetric function of $x_1, \dots, x_k$, and we have:

$$\int_{\mathbb{R}^k} \rho_n^{(k)}(x_1, \dots, x_k) \prod_{i=1}^{k} \mathrm{d}x_i = \frac{n!}{(n-k)!} \,, \tag{43}$$

i.e. the number of ways of choosing k ordered eigenvalues among n. The 1-point Jánossy density coincides with the average spectral density multiplied by n, since

$$\int_{\mathbb{R}} \rho_n^{(1)}(x)\, \mathrm{d}x = n \,.$$

Besides, $\rho_n^{(n)}$ is nothing that the joint p.d.f of the n-eigenvalues, multiplied by $n!$ since (43) gives:

$$\int_{\mathbb{R}^n} \rho_n^{(n)}(x_1, \dots, x_n) = n! \,.$$

The k-point densities can be found by integrating out $(n-k)$ variables in $\rho_n^{(n)}$, again paying attention to the normalization constant:

$$\rho_n^{(k)}(x_1, \dots, x_k) = \frac{1}{(n-k)!} \int_{\mathbb{R}^{n-k}} \rho_n^{(n)}(x_1, \dots, x_n) \prod_{i=k+1}^{n} \mathrm{d}x_i \,. \tag{44}$$

In invariant ensembles

When the random matrix is drawn from an invariant ensemble (Section 8.1), we have:

$$\rho_n^{(n)}(x_1, \dots, x_n) = \frac{n!}{Z_n} \Delta(\lambda_1, \dots, \lambda_n)^2 \prod_{i=1}^{n} e^{-W(\lambda_i)} \,. \tag{45}$$

The Jánossy densities can be computed in terms of the orthogonal polynomials which already appeared in Section 8.2 to compute Z_n. Let us introduce the Christoffel-Darboux kernel:

$$K_n(x, y) = \sum_{k=0}^{n-1} \frac{P_k(x) P_k(y)}{h_k} \,.$$

Using the orthogonality relations, one can easily prove:

$$K_n(x,y) = \frac{P_n(x)P_{n-1}(y) - P_{n-1}(x)P_n(y)}{h_{n-1}(x-y)}\,, \tag{46}$$

which is more advantageous – especially from the point of the large n regime – since it only involves two consecutive orthogonal polynomials.

8.2 theorem.

$$\boxed{\rho_n^{(k)}(x_1,\dots,x_k) = \det_{1\le i,j\le k}\left[\tilde{K}_n(x_i,x_j)\right]}\,, \tag{47}$$

where $\tilde{K}_n(x,y) = K_n(x,y)\,e^{-[W(x_i)+W(x_j)]/2}$.

Proof. We first consider $k = n$. With (45) and Lemma 7.3 and Theorem 8.1, we can write:

$$\rho_n^{(n)}(x_1,\dots,x_n) = \frac{n!}{n!\prod_{m=0}^{n-1} h_m} \det_{1\le i,j\le n}\left[P_{j-1}(\lambda_i)\right] \cdot \det_{1\le k,l\le n}\left[P_{k-1}(\lambda_l)\right] \prod_{i=1}^{n} e^{-W(\lambda_i)}\,.$$

We implicitly used $\det(A^T) = \det(A)$ to write the first determinant. We then push a factor $h_m^{1/2}$ in the columns (resp. in the lines) of the first (resp. the second) determinant, and a factor $\exp[-W(\lambda_m)/2]$ in the lines (resp. the columns) of the first (resp. the second) determinant. The result, using $\det(A\cdot B) = (\det A)\cdot(\det B)$, reads:

$$\begin{aligned}\rho_n^{(n)}(x_1,\dots,x_n) &= \det_{1\le i,j\le n}\left[h_{j-1}^{-1/2}P_{j-1}(\lambda_i)\,e^{-W(\lambda_i)/2}\right] \cdot \det_{1\le k,l\le n}\left[h_{k-1}^{-1/2}P_{k-1}(\lambda_l)\,e^{-W(\lambda_l)/2}\right] \\ &= \det_{1\le i,l\le n}\left[\sum_{k=1}^{n}\frac{P_{k-1}(\lambda_i)P_{k-1}(\lambda_l)}{h_{k-1}}\,e^{-[W(\lambda_i)+W(\lambda_j)]/2}\right],\end{aligned}$$

which is the desired result.

Next, we would like to integrate out the last $n-k$ variables in $\rho_n^{(k)}$ to find $\rho_n^{(k)}$ via (44). This is achieved by successive application of the one-step integration lemma:

8.3 lemma.

$$\int_{\mathbb{R}} \det_{1\le i,j\le k}\left[\tilde{K}_n(x_i,x_j)\right] \mathrm{d}x_k = (n-k+1)\det_{1\le i,j\le k-1}\left[\tilde{K}_n(x_i,x_j)\right]. \tag{48}$$

To prove the lemma, we first remark that $\tilde{K}_n(x,y)$ is the kernel of an operator $\widehat{K}_n : L^2(\mathbb{R},\mathrm{d}x) \longrightarrow L^2(\mathbb{R},\mathrm{d}x)$, which is the orthogonal projection on the

rank n subspace $\mathbb{R}_{n-1}[x] \cdot e^{-W(x)/2}$. In particular – as one can check directly:

$$\begin{aligned} \int_{\mathbb{R}} \tilde{K}_n(x,z)\tilde{K}_n(z,y)\,\mathrm{d}z &= \tilde{K}_n(x,y)\,, \\ \int_{\mathbb{R}} \tilde{K}_n(z,z)\,\mathrm{d}z &= n\,. \end{aligned}$$

Let us expand the $k \times k$ determinant in the left-hand side of (48):

$$\int_{\mathbb{R}} \det_{1\le i,j\le k} \left[\tilde{K}_n(x_i,x_j)\right] \mathrm{d}x_k = \sum_{\sigma\in\mathfrak{S}_k} \mathrm{sgn}(\sigma) \int_{\mathbb{R}} \Big[\prod_{i=1}^{k} \tilde{K}_n(x_i,x_{\sigma(i)})\Big]\,\mathrm{d}x_k\,.$$

We find two types of terms:

- If $\sigma(k) = k$, we have a factor

  $$\int_{\mathbb{R}} \tilde{K}_n(x_k,x_k)\,\mathrm{d}x_k = n.$$

  The remaining factors is a sum over all permutations $\tilde{\sigma} \in \mathfrak{S}_{k-1}$, which reconstructs

  $$\det_{1\le i,j\le k-1} \left[\tilde{K}_n(x_i,x_j)\right].$$

- If $\sigma(k) \neq k$, we rather have a factor

  $$\int_{\mathbb{R}} \tilde{K}_n(x_{\sigma^{-1}(k)},x_k)\,\tilde{K}_n(x_k,x_{\sigma(k)})\,\mathrm{d}x_k = \tilde{K}_n(x_{\sigma^{-1}(k)},x_{\sigma(k)}).$$

  This reconstructs $\prod_{i=1}^{k-1} \tilde{K}_n(x_i,x_{\tilde{\sigma}(i)})$, which only depends on the permutation $\tilde{\sigma} \in \mathfrak{S}_{k-1}$ obtained from σ by "jumping over k", i.e. $\tilde{\sigma}(i) = \sigma(i)$ if $i \neq \sigma^{-1}(k)$, and $\tilde{\sigma}(\sigma^{-1}(k)) = \sigma(k)$. There are exactly $(k-1)$ ways to obtain a given $\tilde{\sigma}$ from some σ, since we have to choose the position of the element $\sigma(k) \in \{1,\dots,k-1\}$. Besides, we have since $\mathrm{sgn}(\tilde{\sigma}) = -\mathrm{sgn}(\sigma)$ since the length of one cycle in $\tilde{\sigma}$ was reduced by 1 compared to σ. All in all, these terms reconstruct:

  $$-(k-1) \det_{1\le i,j\le k-1} \left[\tilde{K}_n(x_i,x_j)\right].$$

Summing the two entails the claim. □

Spectral density

The formula (47) is remarkable: we say that the eigenvalues of hermitian matrices in invariant ensembles form a **determinantal point process**. If $\tilde{K}_n$ were an arbitrary function of two variables, the $k \times k$ determinants of $\tilde{K}_n(x_i,x_j)$ would have no reason to be non-negative. Here, for the Christoffel-Darboux kernel, it must be non-negative by consistency.

For instance, the exact spectral density is $1/n$ times

$$\rho_n^{(1)}(x) = \tilde{K}_n(x,x) = \frac{p_n'(x)p_{n-1}(x) - p_{n-1}'(x)p_n(x)}{h_{n-1}}\, e^{-W(x)}\,. \tag{49}$$

In the GUE

The GUE corresponds to the weight:

$$W(x) = \frac{Nx^2}{2\sigma^2}, \qquad \text{with } N = n\,. \tag{50}$$

We have written N here instead of n, to stress that the size of the matrix appears in two places: first, in the orthogonality weight since W depends on $N = n$, and then in the degree n or $(n-1)$ of the orthogonal polynomials we need to use in (46). To avoid confusion, we may just perform all computations with N, and at the end set $N = n$ to retrieve the GUE normalized as in Section 4. We will also choose $\sigma = 1$.

The orthogonal polynomials for the weight $\mathrm{d}x\, e^{-x^2/2}$ on $\mathbb{R}$ are well-known, called the **Hermite polynomials** and denoted $H_n(x)$. The orthogonal polynomials for the weight $\mathrm{d}x\, e^{-W(x)}$ with (50) are just:

$$P_n(x) = N^{-n/2}\, H_n(N^{1/2}x)\,. \tag{51}$$

We list basic properties of the Hermite polynomials, that can be easily derived using the orthogonality relations:

- H_n has parity $(-1)^n$.
- We have the formula $H_n(x) = (-1)^n\, e^{x^2/2}\, \partial_x^n(e^{-x^2/2})$.
- $H_n'(x) = nH_{n-1}(x)$.
- We have the three-term recurrence relation $H_{n+1}(x) = xH_n(x) - nH_{n-1}(x)$.
- The norm of P_n given by (51) is $h_n = \sqrt{2\pi}\, n!\, N^{-(n+1/2)}$.

Thus, the formula (49) for the spectral density specializes to $1/n$ times (Figure 7):

$$\rho_n^{(1),\mathrm{GUE}}(x) = \frac{\sqrt{n}}{n!\sqrt{2\pi}} \big[H_{n-1}(\sqrt{n}x)\big]^2\, e^{-nx^2/2}\,. \tag{52}$$

In the complex Wishart ensemble

For the Wishart ensemble, one should choose an orthogonality weight on the real positive axis $\mathrm{d}x\, e^{-pV(x)}$ with V given by (39) – and we remind that the size now is denoted p instead of n. The corresponding orthogonal polynomials are also well-known, and called the **Laguerre polynomials**. This makes the computations in the complex Wishart ensemble rather explicit, and amenable to large n asymptotics.

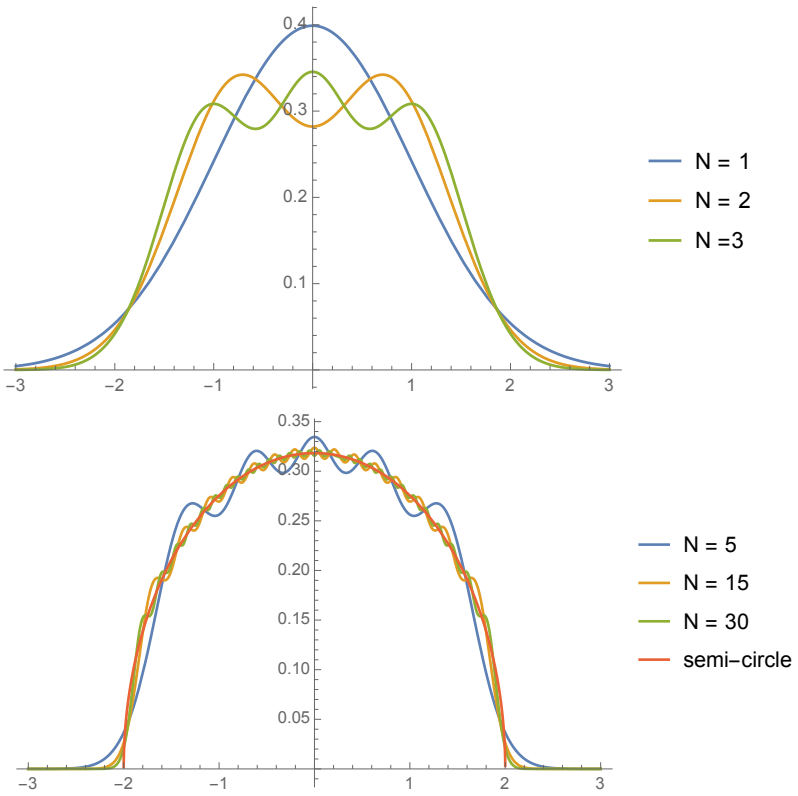


Figure 7: Exact spectral density $n^{-1}\rho_n^{(1)}(x)$ for the GUE with $\sigma = 1$ for matrices of small size n. For $n = 1$, this is just the Gaussian density. For $n \geq 2$ increasing, we see that it approaches the semi-circle law, with oscillations at scale $1/n$. The oscillations for n finite but large can be understood as a consequence of the repulsion of eigenvalues: a region where many eigenvalues are expected prefers having less crowded neighboring regions.

8.4 *Gap probabilities*

The probability that none of the eigenvalues fall into a given measurable set A is also computable in terms of Jánossy densities:

$$\begin{aligned}
\mathbb{P}\big[\text{no eigenvalue in } A\big] &= \mathbb{E}\Big[\prod_{i=1}^{n}(1-\mathbf{1}_A(\lambda_i))\Big] \\
&= \sum_{k=0}^{n}(-1)^k \sum_{1\leq i_1<\ldots<i_k\leq n} \mathbb{P}\big[\lambda_{i_1},\ldots,\lambda_{i_k}\in A\big] \\
&= \sum_{k=0}^{n}\frac{(-1)^k}{k!}\int_{A^k}\rho_n^{(k)}(x_1,\ldots,x_k)\prod_{i=1}^{k}\mathrm{d}x_i\,. \qquad (53)
\end{aligned}$$

From (47), we find:

$$\mathbb{P}\big[\text{no eigenvalues in } A\big] = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \int_{A^k} \det_{1\le i,j\le k} \big[\tilde{K}_n(x_i, x_j)\big] \prod_{i=1}^{k} \mathrm{d}x_i \,.$$

Since K_n is the kernel of an operator of rank n, the determinants of size $k > n$ vanish, and we have:

$$\mathbb{P}\big[\text{no eigenvalues in } A\big] = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \int_{A^k} \det_{1\le i,j\le k} \big[\tilde{K}_n(x_i, x_j)\big] \prod_{i=1}^{k} \mathrm{d}x_i \,.$$

We recognize the definition of the Fredholm determinant[12] of the operator $\widehat{K}_n$ restricted to act on the Hilbert space $L^2(A, \mathrm{d}x)$:

$$\mathbb{P}\big[\text{no eigenvalues in } A\big] = \mathrm{Det}\left[1 - \widehat{K}_n\right]_{L^2(A,\mathrm{d}x)} . \tag{54}$$

The Fredholm determinant $\mathrm{Det}[1 - \widehat{K}]$ is a continuous function of $\widehat{K}$ for the topology induced by the sup-norm for the kernel $\widehat{K}(x, y)$ of $\widehat{K}$. This means that, to study the large n asymptotics of (54), it is enough to study the uniform convergence of the kernel $\tilde{K}_n(x, y)$.

In particular, if we take A to be the semi-infinite interval $(a, +\infty)$, the probability that no eigenvalue belongs to A is exactly the probability that the maximum eigenvalue is smaller than a:

$$\mathbb{P}[\lambda_{\max} \le a] = \mathrm{Det}\left[1 - \widehat{K}_n\right]_{L^2\left((a,+\infty),\mathrm{d}x\right)} .$$

[12] This is a generalization of the notion of determinant to operators in infinite-dimensional spaces.

9 Asymptotics and universality of local regime

We have expressed the Jánossy densities and the gap probabilities in terms of the Christoffel-Darboux kernel:

$$\tilde{K}_n(x,y) = \frac{P_n(x)P_{n-1}(y) - P_{n-1}(x)P_n(y)}{h_{n-1}(x-y)} \,. \tag{55}$$

In order to study the large n limit of the eigenvalue distributions, we just need to derive the asymptotics of the orthogonal polynomials $P_n(x)$.

9.1 *Asymptotics of Hermite polynomials*

For Hermite polynomials, one can easily establish, from the properties previously mentioned, the integral representation:

$$H_n(x) = \mathrm{i}^n \, e^{x^2/2} \int_{\mathbb{R}} \mathrm{d}\zeta \, \zeta^n \, e^{-\zeta^2/2 - \mathrm{i}x\zeta} \,.$$

The asymptotics of $H_n(x)$ can then be derived using the classical method of steepest descent analysis[13] – see e.g. [1] for details. The result is called the Plancherel-Rotach formula – see e.g. [30]. Let us define:

$$\varphi_n(x) = \frac{e^{-x^2/4} \, H_n(x)}{\sqrt{\sqrt{2\pi}\, n!}} \,.$$

9.1 theorem. *Let m be a fixed integer, and consider $n \to \infty$.*

- ***Bulk.*** *For fixed $x_0 \in (-2,2)$ and $X \in \mathbb{R}$, we have:*

$$\varphi_{n+m}(n^{1/2}x_0 + n^{-1/2}X) = \frac{2\cos\left[\theta_n(x_0,X,m)\right]}{n^{1/4}\sqrt{2\pi}(4-x_0^2)^{1/4}} + O(n^{-3/4}) \,, \tag{56}$$

with:

$$\begin{aligned} \theta_n(x_0,X,m) &= (n+m+1)\arcsin(x_0/2) - \frac{\pi(n+m)}{2} \\ &\quad + \frac{nx_0\sqrt{4-x_0^2}}{4} + \frac{X\sqrt{4-x_0^2}}{2} \,. \end{aligned}$$

The result is uniform for X in any compact of $\mathbb{R}$.

- ***Edge.*** *For fixed $X \in \mathbb{R}$, we have:*

$$\varphi_{n+m}(2n^{1/2} + n^{-1/6}X) = n^{-1/12}\,\mathrm{Ai}(X) + O(n^{-5/12}) \,, \tag{57}$$

where Ai *is the Airy function, i.e. the unique solution to* $\mathrm{Ai}''(X) = X\mathrm{Ai}(X)$

[13] This is a generalization in complex analysis of the Laplace method in real analysis to study the $\epsilon \to 0$ behavior integrals of the form $\int_{\mathbb{R}} e^{-f(x)/\epsilon})\mathrm{d}x$.

which decays[14] *when* $X \to +\infty$ *like:*

$$\mathrm{Ai}(X) \sim \frac{\exp\left(-\frac{2}{3}X^{3/2}\right)}{2\sqrt{\pi}\,X^{1/4}}\,.$$

(57) *is uniform for* X *in any compact of* $\mathbb{R} \cup \{+\infty\}$.

- ***Far side.*** *For fixed* $|x_0| > 2$, $\varphi_{n+m}(n^{1/2}x_0)$ *decays exponentially fast when* $n \to \infty$.

The existence of the three regimes have direct qualitative consequences for the distribution of eigenvalues in the large n limit. In the bulk, the Hermite polynomials have an oscillatory asymptotics: it is the region where their n zeroes accumulate, and where the eigenvalue distribution will be concentrated. As expected, with the scaling (51), we look at arguments of the Hermite polynomials at the scale $\sqrt{n}$, and the bulk thus correspond to the bounded interval $x_0 \in [-2, 2]$. In (56), we see that non-trivial variations occur when we deviate from x_0 with order of magnitude $1/n$, as measured by X. This means that fluctuations of eigenvalues in the bulk of the GUE will occur at scale $O(1/n)$. The result in the far side indicates that it will be exponentially unlikely to find eigenvalues outside of $[-2, 2]$, and confirms that the support of the spectral density should be $[-2, 2]$. At the right edge $x_0 = 2$ between the far side and the bulk – the behavior at the left edge $x_0 = -2$ is obtained by symmetry – there is a transition, and non-trivial variations now occur when x_0 deviates from 2 with order of magnitude $n^{-1/2} \cdot n^{-1/6} = n^{-2/3}$. So, the fluctuation of eigenvalues near the edge, and in particular the fluctuations of the maximum, will be of order $n^{-2/3}$, as anticipated in Section 4.2.

Notice that the introduction of the variable X in Theorem 9.1 allows to reach the distribution of eigenvalues in regions where only finitely many eigenvalues are expected – these are regions of size $1/n$ in the bulk, and of size $n^{-2/3}$ around the edge – i.e. the local regime, while keeping only x_0 would provide information about the global regime only.

There is no difficulty in computing the asymptotics of the Christoffel-Darboux kernel (55) in the various regimes from Theorem 9.1, although the algebra is a bit lengthy. We shall summarize the results of these computations.

9.2 *Consequences in the bulk*

First, we find that the spectral density converges to the semi-circle law:

$$\lim_{n\to\infty} n^{-1}\rho_n^{(1)}(x_0) = \frac{\sqrt{4-x_0^2}}{2\pi}\,\mathbf{1}_{[-2,2]}(x_0)\,.$$

[14] At $X \to -\infty$, $\mathrm{Ai}(X)$ is unbounded and has oscillatory asymptotics.

For the local regime around a point $x_0 \in (-2,2)$ in the bulk, we find:

$$\lim_{n\to\infty} \frac{\tilde{K}_n\left(x_0 + \frac{X}{\rho_n^{(1)}(x_0)}, x_0 + \frac{Y}{\rho_n^{(1)}(x_0)}\right)}{\rho_n^{(1)}(x_0)} = \frac{\sin \pi(X-Y)}{\pi(X-Y)}\,. \tag{58}$$

This function is called the **sine kernel**, and denoted $K_{\sin}(X,Y)$. The corresponding operator is denoted $\widehat{K}_{\sin}$. In (58), It was natural, instead of choosing to measure X in units of $1/n$, to normalize it further by the spectral density. Indeed, the average local density of eigenvalues measured in terms of X is equal to 1, and this facilitates the comparison between different models.

9.2 corollary. *For any fixed integer k, and fixed $x_0 \in (-2,2)$, the eigenvalue distribution is such that:*

$$\lim_{n\to\infty} \frac{\rho_n^{(k)}\left[\left(x_0 + \frac{X_i}{\rho_n^{(1)}(x_0)}\right)_{i=1}^n\right]}{\rho_n^{(1)}(x_0)^k} = \det_{1\le i,j\le k} K_{\sin}(X_i,X_j)\,.$$

And, for any compact A of $\mathbb{R}$, the gap probability behaves like:

$$\lim_{n\to\infty} \mathbb{P}\left[\text{no eigenvalue in } \left(x_0 + \frac{A}{\rho_n^{(1)}(x_0)}\right)\right] = \mathrm{Det}\left[1-\widehat{K}_{\sin}\right]_{L^2(A,\mathrm{d}x)},$$

where $a + b\cdot A$ the image of A by the map $x \mapsto a + bx$.

9.3 *Consequences at the edge*

We find that the Christoffel-Darboux kernel at the edge behaves like:

$$\lim_{n\to\infty} n^{-1/6}\,\tilde{K}_n(2+n^{-2/3}X, 2+n^{-2/3}Y) = \frac{\mathrm{Ai}(X)\mathrm{Ai}'(Y) - \mathrm{Ai}'(X)\mathrm{Ai}(Y)}{X-Y}\,.$$

This is the **Airy kernel**, denote $K_{\mathrm{Ai}}(X,Y)$. The corresponding operator is denoted $\widehat{K}_{\mathrm{Ai}}$.

9.3 corollary. *At the right edge of the spectrum, the eigenvalue distribution is such that:*

$$\lim_{n\to\infty} n^{-k/6}\,\rho_n^{(k)}\left[(2+n^{-2/3}X_i)_{i=1}^n\right] = \det_{1\le i,j\le k} K_{\mathrm{Ai}}(X_i,X_j)\,.$$

And, for any compact A of $\mathbb{R}\cup\{+\infty\}$, the gap probability behaves like:

$$\lim_{n\to\infty} \mathbb{P}\left[\text{no eigenvalue in } 2+n^{-2/3}A\right] = \mathrm{Det}[1-\widehat{K}_{\mathrm{Ai}}]_{L^2(A,\mathrm{d}x)}\,.$$

In particular:

$$\lim_{n\to\infty} \mathbb{P}[\lambda_{\max} \le 2+n^{-2/3}s] = \mathrm{Det}[1-\widehat{K}_{\mathrm{Ai}}]_{L^2\left((s,+\infty),\mathrm{d}x\right)}\,.$$

is another expression – the first historically obtained – of the Tracy-Widom law $\mathrm{TW}_2(s)$.

9.4 *Universality*

Here is a table summarizing the limit distributions we have seen.

Jánossy densities	gap probabilities	universality class
$\det_{k\times k}\left[K_{\sin}(x_i,x_j)\right]$	$\mathrm{Det}\left[1-\widehat{K}_{\sin}\right]_{L^2(A)}$	in bulk $\rho(x_0)\sim$ cte
$\det_{k\times k}\left[K_{\mathrm{Ai}}(x_i,x_j)\right]$	$\mathrm{Det}\left[1-\widehat{K}_{\mathrm{Ai}}\right]_{L^2(A)}$	at edge $\rho(x_0)\sim(a-x_0)^{1/2}$

They are universal – i.e. valid independently of the details of the model – for hermitian random matrices in invariant ensembles, for complex Wishart matrices, and many other ensembles of random hermitian matrices. For symmetric matrices, there exist different universal laws – we have seen an expression of $\mathrm{TW}_1(s)$ – which are also well understood [26]. This universality goes actually beyond random matrices, see e.g. the review [10]. Let us illustrate it by two examples.

Non-intersecting random walks

Consider the standard brownian motion (BM) in $\mathbb{R}$, and let $\mathcal{K}_t(x,y)$ be the probability density that a BM starting at time $t=0$ at position x, ends at time t at position y. It is a basic result of stochastic processes that:

$$\mathcal{K}_t(x,y)=(2\pi t)^{-1/2}\exp\Big(-\frac{(x-y)^2}{2t}\Big).$$

Since BM is a Markov process, we also have:

$$\int_{\mathbb{R}}\mathcal{K}_t(x,z)\,\mathcal{K}_{t'}(z,y)\,\mathrm{d}z=\mathcal{K}_{t+t'}(x,y).$$

Now, let us consider n independent BMs starting from positions $x_1<\ldots<x_n$ at time $t=0$, and conditioned not to intersect. Karlin and McGregor in 1960 [23] have computed the probability density that they arrive at time t at positions $y_1<\ldots<y_n$:

$$\mathcal{P}_n(x_1,\ldots,x_n|y_1,\ldots,y_n)=\det_{1\le i,j\le n}\mathcal{K}_t(x_i,y_j)$$

This is the starting point of a series of results, showing that in various situations, the non-intersecting random walkers – sometimes called vicious because they do not want to cross – behave when $n\to\infty$ like eigenvalues of large random matrices (Figure 8). For instance, the fluctuations of the position of the rightmost walker generically occur at scale $n^{-2/3}$ around their mean, and converge in law towards the Tracy-Widom GUE law. Similarly, if one zoom amidst the walkers in a region where we expect to see only finitely many of them, the distribution of the positions of k of them is given by the

$k \times k$ determinant build from the sine kernel. More details can be found in [12].

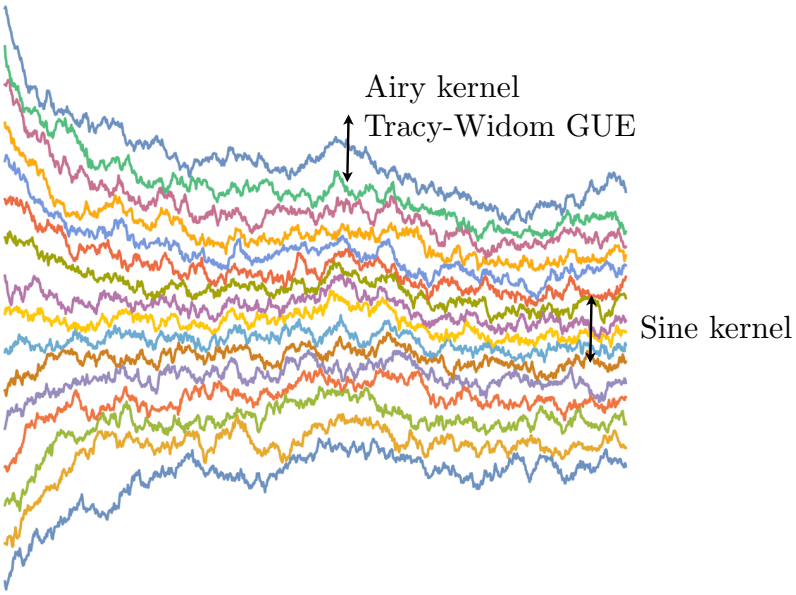


Figure 8: Simulation (courtesy of P. Ferrari) n independent random walks in 1 dimension, conditioned not to intersect. In the large n limit, after proper rescaling, the fluctuations of the height of the top path follows the Tracy-Widom GUE law, and the joint distribution of a finite number of paths starting from the top path is given by the determinantal process with kernel K_{Ai}. For a path in the bulk, the fluctuations of the height of a finite number of consecutive paths are given by the determinantal process with kernel $K_{\sin}$.

Growth models

The sine kernel or the Airy kernel distributions also appear in problems of growing interfaces. There exist several mathematical models where this has been established – see the review [13]. But I also want to point out, with an example, that these distributions can be seen in (even non-mathematical) nature.

The physicists Takeuchi and Sano (2010) observed experimentally the Tracy-Widom law in nematic liquid crystals. "Nematic" means that the material is made of long molecules whose orientation has long-range correlations, while liquid means that the molecules in the neighborhood of a given one are always changing, i.e. the correlation of positions have short range. In nematic materials, a "topological defect" is a configuration of orientations that winds around a point. In 2d, it occurs for instance when the local orientation rotates like the tangent vector when following a circle throughout the material[15]. The mate-

[15] In three dimensions, the Hopf fibration $\phi : S_3 \to S_2$ is a configuration of orientations realizing

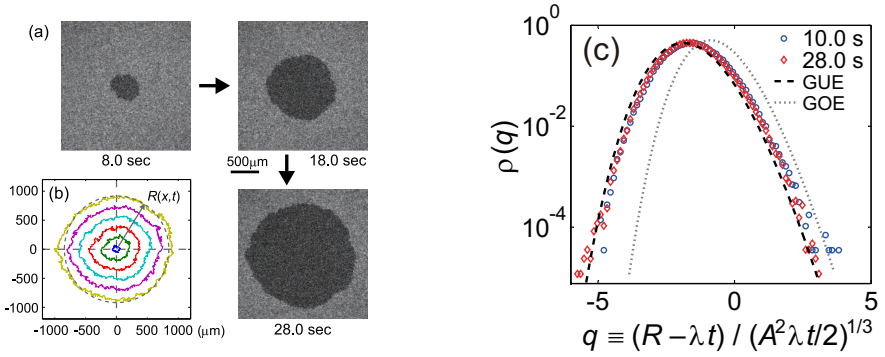


Figure 9: Comparison between fluctuations of the radius of a growing interface in nematic liquid crystals and Tracy-Widom laws. Reprinted with permission from *Universal fluctuations of growing interfaces: evidence in turbulent liquid crystals*, K. Takeuchi and M. Sano, Phys. Rev. Lett. **104** 230601 (2010) © APS.

rial studied by Takeuchi and Sano presents two phases: the phase appearing here in gray (resp. black) has a low (resp. high) density of topological defects. If one applies a voltage to the grey phase, one encourages the formation of defects. Once this happens – here at the center of the picture at time $t = 0$ – the black phase takes over the grey phase from this primary cluster of defects. One observes that the interface grows approximately linearly with time t. However, the turbulence driving the system causes some fluctuations from samples to samples. The distribution of these fluctuations of radius from the linear drift matches with the Tracy-Widom GUE law, and the quality of the fit improves with time increasing (Figure 9). The symmetry class in this case is conditioned by the geometry: a spherical geometry leads to GUE, while a flat interface between two phases would lead to GOE. This result is confirmed in a mathematical model for the interface growth analyzed at $t \to +\infty$ by Sasamoto and Spohn around the same time [29].

Last remarks

In the last 20 years, tremendous progress has been made to prove universality in random matrices, with weak assumptions, relying on various approaches. Without exhaustivity, we can cite:

- the fact that some models are exactly solvable (like the invariant ensembles of symmetric or hermitian random matrices) and Riemann-Hilbert steepest descent analysis. This is very useful, but maybe not very satisfactory from the probabilistic point of view, since the method hinges from the beginning on "algebraic miracles", that disappear if the models are slightly perturbed.

a topological defect.

- transport of measures (Shcherbina ; Figalli, Guionnet and Bekerman), which has succeeded in proving some universality for all β-ensembles.
- relaxation methods (Bourgade, Erdös, H.T.-Yau, etc.) which are of purely based on probability, stochastic processes and analysis, which brought many results for invariant ensembles, matrices with independent entries, etc.
- combinatorial methods (Wigner ; Soshnikov ; Tao and Vu, etc.) that are particularly useful for matrices with independent entries, etc.

One current trend is now to apply the insight gained from the study of random matrices, to more difficult problems like random band matrices, random Schrödinger operators, adjacency matrices of random graphs, etc. This is motivated by the desire to understand the properties of localization/delocalization of the eigenvectors – that determine isolating/conducting properties of materials modelized in this way.

10 Questions of participants

• Ninjbat Uuganbaatar: Can one apply PCA techniques to analyze voting ?

In general, the number of options for which one votes is very small, so I do not see how PCA can be used to analyze voting. However, it could be a tool to check the representativity of the political offer in a given society. For instance, one could ask n individuals to answer a poll consisting of p questions about their political preference. As example of questions: how much should income be taxed ? at which age should people retire ? should the state subsidize health coverage ? ... The opinion pollster would have to choose a way to get answers which are numbers, for instance binary questions – somewhat like in population genetics about presence or absence of an allele – given 0 or 1 as entries, or questions that one can answer by an intensity from 0 (not at all) to 10 (absolutely). Then, one can build a $n \times p$ matrix X collecting the answers, and the empirical covariance matrix $M = p^{-1} XX^T$. By PCA analysis, one can then hope to determine how many relevant groups can be formed, that have similar political ideas – as measured by the questions asked. One could then compare with the number of political parties, as well as their programme, to see if the population is well-represented at the level of ideas, and if their strength compares well with the magnitude of the eigenvalues found in PCA. I do not know if such a project has been already conducted. Clearly, an important work of calibration is needed – e.g. checking if the outcome of PCA is similar when one asks yes/no questions, or intensity questions, etc. – to ensure the results are reliable.

• Remco van der Hofstad: How can one identify quantitatively in PCA what comes from true information and what comes from noise ?

For market prices, we have seen in the examples of Section 3.4 that the overlap between the j-th eigenvector – sorted by decreasing order for the corresponding eigenvalues – of empirical correlation matrices in two distinct periods does not exceed what one expects from the overlap of two independent random vectors for some $j \geq j_0$. And this threshold also corresponded well with the position of the noise band – i.e. the distribution of eigenvalues λ_j with $j \geq j_0$ was fitted with the Marčenko-Pastur law.

A more general method is to fix a confidence threshold, and then make a statistical test for λ_i using the Tracy-Widom law, for $i = 1, 2, 3, \ldots$ until one cannot reject anymore the null hypothesis (which enjoys Tracy-Widom distribution). More precisely, if the test is passed for λ_i, one restricts the matrix to the orthogonal of the eigenspace of $\lambda_1, \ldots, \lambda_i$ before continuing the analysis. And there exists estimates of the rate of convergence to the Tracy-Widom law in null Wishart matrices (see e.g. [22]) when n, p is large but not infinite, which can be used for statistical tests. To cope with finite size effects, one can also use large deviation functions – see the question below – but one should keep in mind that their details are much less robust (if one changes the model) than the Tracy-Widom distribution.

• Kanstantsin Matetski: What can be said about the large deviations of the maximum eigenvalue ?

Although I did not present them for lack of space in the lectures, there exist techniques, based on potential theory and large deviation theory, to compute the asymptotic behavior of the partition function in invariant ensembles. In particular, if one assume that the support of the large n spectral density is a single segment (as for GUE and Wishart) + some other technical assumptions on V, one can show that the partition function:

$$Z_{n,\beta}(A) = \int_{A^n} \prod_{1\leq i<j\leq n} |\lambda_j - \lambda_i|^\beta \prod_{i=1}^n \exp\Big(-\frac{n\beta}{2} V(\lambda_i)\Big) \mathrm{d}\lambda_i$$

has an asymptotic expansion of the form:

(59)
$$\ln Z_{n,\beta}(A) = n^2 F_0 + (\beta/2) n \ln n + n(\beta/2-1) F_1 + \frac{3+2/\beta+\beta/2}{12} \ln n + F_2 + o(1)$$

when $n \to \infty$, and the coefficients F_j can be computed fairly explicitly, depending on V and A. The $o(1)$ actually consists of a full asymptotic expansion in powers of $1/n$, and its coefficients can also be computed recursively.

These results give access to the large deviations for the maximum eigenvalue, since:

$$\mathbb{P}[\lambda_{\max} \leq a] = \frac{Z_{n,\beta}(a, +\infty)}{Z_{n,\beta}(\mathbb{R})}.$$

For instance, when a is independent of n and strictly smaller than $a_* = \lim_{n\to\infty} \mathbb{E}[\lambda_{\max}]$, the assumptions leading to (59) are satisfied and we can prove rigorously an asymptotic expansion of the form:

(60)
$$\mathbb{P}[\lambda_{\max} \leq a] = n^c \exp\Big[-n^2 G_0(a) - n(\beta/2-1) G_1(a) + \sum_{k\geq 0}^K n^{-k} G_{k+2}(a) + o(n^{-K})\Big].$$

For $a < a_*$, this probability is super-exponentially small because one has to push all the n eigenvalues to the left of a_* to achieve the event $\lambda_{\max} \leq a < a_*$. The leading term $G_0(a)$ is called the large deviation function, and has some relevance in statistical applications, because one has to face the finite size of data.

How does that connect to the Tracy-Widom law ? If one naively insert $a = a_* - s n^{-2/3}$ in the right-hand side of (60), we can show that each term $n^{-k} G_{k+2}(a)$ tends to a constant $\tilde{G}_{k+2}\, s^{-3k/2}$, which is of order 1. This is not surprising because in this regime the probability (60) should vary between 0 and 1. As a matter of fact, putting $a = a_* - s n^{-2/3}$ goes out of the range in which (60) was established. But, if one is ready to believe that the crossover from "large deviations" to "not so large deviations" is smooth – an exchange of limits that has not been justified as of writing – then we interpret the naive

right-hand side where one first inserts $a = a^* - sn^{-2/3}$ as the all-order asymptotic expansion when $s \to +\infty$ of $\mathrm{TW}_\beta(-s)$. This leads to predictions, for any value of $\beta > 0$, for the left tail of Tracy-Widom β laws. They agree with all rigorous results known for $\beta = 1, 2$, and with the leading order rigorously known for arbitrary β. In particular, we have a prediction for the constant term of the asymptotic expansion, which is always tricky to get. A similar story can be devised for the right tail.

The large deviation function $G_0(a)$ at the left tail was first computed by Dean and Majumdar in [8] – although this is a physics paper, the equation they solve to get $G_0(a)$ can be rigorously established using potential theory without any difficulty, hence making a complete proof. We discussed the generalization to all-order finite size corrections in [4] for the left tail, and [6] for the right tail. The computations in these two papers are done for the Gaussian ensembles, but there would be not difficulty in conducting them for other V, e.g. for the Wishart ensembles. These two papers take as starting point the asymptotic expansion of the form (60) ; these expansions have been established rigorously in [5].

• Ninjbat Uuganbaatar : Is there a combinatorial interpretation to the formulas we have seen for the distribution of random matrices ?

Let us start with a matrix M_n in the Gaussian ensembles, for $\sigma = 1$. The moments of the semi-circle law can be directly computed by expanding its Stieltjes transform (14) at $z \to \infty$:

$$\lim_{n\to\infty} n^{-1} \operatorname{Tr} M_n^{2k} = \frac{2k!}{k!(k+1)!} = \mathrm{Cat}(k)\,.$$

This is the Catalan number, computing the number of ways to connect pairs of edges in a $2k$-gon, without crossing. More generally, Harer and Zagier in 1986 [18] showed the expansion:

$$\operatorname{Tr} M_n^{2k} = \sum_{g\geq 0} n^{1-2g}\, \mathcal{N}_n(g)$$

where $\mathcal{N}_n(g)$ is the number of ways of identifying by pairs the edges of $2k$-gon, in such a way that the resulting surface has genus g. They gave several formulas to compute these numbers – from (52), we know that they can be expressed in terms of Hermite polynomials. Harer and Zagier used this to compute the Euler characteristics of the moduli space of Riemann surfaces of genus g ; this is one of the many and fruitful point of contacts between random matrices and algebraic geometry.

Actually, the combinatorial interpretation of the moments of the GUE was already known to physicists, in the more general context of invariant ensembles of hermitian matrices. Brézin, Itzykson, Parisi and Zuber showed in 1979 [7] that the partition function "decomposes" as:

$$Z_n = n^{n+5/12} \exp\Big(\sum_{g\geq 0} n^{2-2g}\, \mathcal{F}_g \Big)\,,$$

and $\mathcal{F}_g$ enumerates discretized surfaces of genus g. For instance, if one takes $V(x) = x^2/2 - tx^3/3$, $\mathcal{F}_g$ is the number of triangulations of a genus g surface, counted with a weight t^T if it is made exactly of T triangles. Although it seems he partition function does not make sense as a convergent integral since $V(x) \to -\infty$, it can be defined rigorously as a formal series in the parameter t – and this is why I said "decompose" with quotes. Likewise the expectation values:

$$\mathbb{E}\left[\operatorname{Tr} M_n^{\ell_1} \cdots \operatorname{Tr} M_n^{\ell_k}\right]$$

are related to the enumeration of discretized surfaces with k boundaries of respective perimeters $\ell_1, \ldots, \ell_k$ counted with a weight n^χ where χ is the Euler characteristics. The coupling of the matrix size with the Euler characteristics is a phenomenon that was first observed in gauge theories by the theoretical physicist t'Hooft in 1974 [32]. More on the relations between random matrices, enumeration of discretized surfaces and algebraic geometry, can be found in the book [11].

References

[1] G.W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, Cambridge University Press, 2010.

[2] J. Baik, G. Ben Arous, and S. Péché, *Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices*, Ann. Probab. **33** (2005), 1643–1697, math.PR/0403022.

[3] A. Bloemendal and B. Virág, *Limits of spiked random matrices I*, PTRF **156** (2013), no. 3-4, 795–825, math.PR/1011.1877.

[4] G. Borot, B. Eynard, S.N. Majumdar, and C. Nadal, *Large deviations of the maximal eigenvalue of random matrices*, J. Stat. Mech. (2011), no. P11024, math-ph/1009.1945.

[5] G. Borot and A. Guionnet, *Asymptotic expansion of β matrix models in the one-cut regime*, Commun. Math. Phys **317** (2013), no. 2, 447–483, math.PR/1107.1167.

[6] G. Borot and C. Nadal, *Right tail expansion of Tracy-Widom beta laws*, RMTA **1** (2012), no. 03, math-ph/1111.2761.

[7] É. Brézin, C. Itzykson, G. Parisi, and J.-B. Zuber, *Planar diagrams*, Commun. Math. Phys. **59** (1978), 35–51.

[8] D.S. Dean and S.N. Majumdar, *Large deviations of extreme eigenvalues of random matrices*, Phys. Rev. Lett. **97** (2006), 160–201, cond-mat/0609651.

[9] P. Deift, *Orthogonal polynomials and random matrices : a Riemann-Hilbert approach*, AMS, New York, 1998, Courant Institute of Mathematical Sciences.

[10] ———, *Universality for mathematical and physical systems*, Proceeding of the ICM, Madrid 2006, Spain (2007), 125–152, math.ph/0603038.

[11] B. Eynard, *Counting surfaces*, Progress in Mathematics, Birkhäuser, 2016, available at http://eynard.bertrand.voila.net/TOCbook.htm.

[12] P.L. Ferrari, *Why random matrices share universal processes with interacting particle systems ?*, (2013), ICTP Lecture notes, math-ph/1312.1126.

[13] P.L. Ferrari and H. Spohn, *Random growth models*, (2010), math.PR/1003.0881.

[14] R.A. Fisher, *The sampling distribution of some statistics obtained from non-linear equations*, Ann. Eugenics **9** (1939), 238–249.

[15] P.J. Forrester, *The spectrum edge of random matrix ensembles*, Nucl. Phys. B (1993), 709–728.

[16] S. Geman, *A limit theorem for the norm of random matrices*, Ann. Probab. **8** (1980), no. 2, 252–261.

[17] M.A. Girshick, *On the sampling theory of roots of determinantal equations*, Ann. Math. Stat. **10** (1939), 203–204.

[18] J. Harer and D. Zagier, *The Euler characteristics of the moduli space of curves*, Invent. Math. **85** (1986), 457–485.

[19] S.P. Hastings and J.B. McLeod, *A boundary value problem associated with the second Painlevé transcendent and the Korteweg-de Vries equation*, Archive for Rational Mechanics and Analysis **73** (1980), no. 1, 31–51.

[20] H. Hotelling, *Analysis of a complex of statistical variables into its principal components*, Journal of Educational Psychology (1931), 417–441.

[21] P.L. Hsu, *On the distribution of roots of certain determinantal equations*, Ann. Eugenics **9** (1939), 250–258.

[22] I.M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Ann. Stat. **29** (2001), no. 2, 295–327.

[23] S. Karlin and J. McGregor, *Coincidence probabilities*, Pacific J. Math. **9** (1959), no. 4, 1141–1164.

[24] H. Markowitz, *Portfolio selection*, J. Finance **7** (1952), no. 1, 77–91.

[25] V.A. Marčenko and L.A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Mat. Sb. **72** (1967), no. 4, 507–536.

[26] M.L. Mehta, *Random matrices*, 3$^{\text{ème}}$ ed., Pure and Applied Mathematics, vol. 142, Elsevier/Academic, Amsterdam, 2004.

[27] K. Pearson, *On lines and planes of closest fit to systems of points in space*, Philosophical Magazine **2** (1901), 559–572.

[28] S.N. Roy, *p-statistics or some generalizations in the analysis of variance appropriate to multivariate problems*, Sankhya **4** (1939), 381–396.

[29] H. Spohn and T. Sasamoto, *The one-dimensional KPZ equation: an exact solution and its universality*, Phys. Rev. Lett. **104** (2010), cond-mat.stat-mech/1009.1883.

[30] G. Szegö, *Orthogonal polynomials*, Amer. Math. Soc., 1939, reprinted with corrections (2003).

[31] T. Tao, *Topics in random matrix theory*, Graduate Studies in Mathematics, vol. 132, AMS, 2012.

[32] G. t'Hooft, *A planar diagram theory for strong interactions*, Nucl. Phys. B **72** (1974), 461–473.

[33] C. Tracy and H. Widom, *Level spacing distributions and the Airy kernel*, Commun. Math. Phys. **159** (1994), 151–174, hep-th/9211141.

[34] ———, *On orthogonal and symplectic matrix ensembles*, Commun. Math. Phys. **177** (1996), 727–754, solv-int/9509007.

[35] D.V. Voiculescu, *Limit laws for random matrices and free products*, Invent. Math. **104** (1991), 201–220.

[36] E.P. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. Math. **62** (1955), no. 3, 548–564.

[37] J. Wishart, *The generalised product moment distribution in samples from a normal multivariate population*, Biometrika **20A** (1928), no. 1/2, 32–52.