The loss of matrix norm equivalence in big data analysis and the Marchenko-Pastur Law

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial pulfillment of the requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By
Emma Heidorn

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics & Statistics
142 McLean Hall
106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

OR

Dean

College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

In statistics, p dimensional data are collected n times. Traditionally, the dimension of p would be larger than n; however, as technology progresses, we enter the era of big data where n is no longer much larger than p. The large ratio of $\frac{p}{n}$ causes pitfalls in methods and algorithms that were developed with the opposite in mind. To solve this problem, methods using random matrix theory were brought up in [4], this thesis will be focusing on results concerning the Marchenko-Pastur Law.

This thesis is not a cutting-edge research, but an organized presentation of the Marchenko-Pastur Law. This is written so students and researchers can quickly grasp the ideas and methods without difficulty.

Contents

Introduction

The main objective of this thesis is to study the Marchenko-Pastur Law. In 1967, Ukrainian mathematicians V.A. Marchenko and L.A. Pastur proved the Marchenko-Pastur Law, which describes the asymptotic behavior of eigenvalues of large sample covariance random matrices in their paper [6]. This thesis is not a cutting-edge research, but instead, provides a detailed proof for people studying the Marchenko-Pastur Law to understand the theorem and its proof.

Our proof of the Marchenko-Pastur Law in Chapter 2 roughly follows the one for Wigner's Semicircle Law in the book [7]. In fact, our approach has been listed as a series of exercises in [7], however, the complete solutions to these exercises are not provided in [7].

This thesis is organized into three chapters:

- Chapter 1 introduces definitions involving random matrices and the problem of the loss of matrix norm equivalence that we encounter in large rectangular matrices, which motivates our use of the Marchenko-Pastur Law.

- Chapter 2 introduces the Marchenko-Pastur Law. Before proceeding to the proof, further definitions and theorems are introduced to be used. The proof is further broken up into three sections: convergence in moments, limit distribution, and convergence in distribution.

- Chapter 3 concludes by elaborating the applications of the Marchenko-Pastur Law and additional resources for further reading.

1 The loss of equivalence between matrix norms

1.1 Motivating the Marchenko-Pastur Law

We begin by introducing the matrix norm and its properties.

Definition (Matrix Norm). The matrix norm is the norm on the space of $N \times N$ complex or real matrices ($M_N(\mathbb{C})$ or $M_N(\mathbb{R})$) denoted by $||\cdot||$ that satisfies the following:

- $||A|| \geq 0$.
- $||A|| = 0 \rightarrow A = 0 \in M_N(\mathbb{C})$ or $M_N(\mathbb{R})$.
- $||A + B|| \leq ||A|| + ||B||$ (Triangle inequality).

Example 1.1.1. Let $A = [a_{ij}]_{N\times N}, a_{ij} \in \mathbb{C}$. Taking the 1-norm, we have

$$||A||_1 = \sum_{ij} |a_{ij}|.$$

Taking the p-norm, $1 \leq p \leq \infty$, we have

$$||A||_p = \left(\sum_{ij} |a_{ij}|^p \right)^{1/p}.$$

Taking the infinity norm, we have

$$||A||_\infty = \max_{i,j} |a_{ij}|.$$

Taking the operator norm of we have

$$||A|| = \max \left\{ |Ax| : x \in \mathbb{C}^N, \text{with } |x| = 1 \right\}.$$

Similarly, taking the operator norm of $A \in M_N(\mathbb{R})$, we have

$$\|A\| = \max\left\{|Ax| : x \in \mathbb{R}^N \text{ with } |x| = 1\right\}.$$

Proposition 1.1.1. (matrix norm equivalence on finite dimension). If $||\cdot||^{(1)}$ and $||\cdot||^{(2)}$ are two matrix norms on $M_N(\mathbb{C})$, then there exist constants $c_1, c_2 > 0$ such that

$$c_1||A||^{(1)} \leq ||A||^{(2)} \leq c_2||A||^{(1)} \quad \forall A \in M_N(\mathbb{C}).$$

Proof. Given $A \in M_N(\mathbb{C})$ with $A \neq 0$, one has

$$\begin{aligned}
||A||^{(2)} &= ||A||^{(1)} \cdot \frac{||A||^{(2)}}{||A||^{(1)}} \\
&= ||A||^{(1)} \cdot \left\| \frac{A}{||A||^{(1)}} \right\|^{(2)} \\
&\leq ||A||^{(1)} \cdot \sup_{B \neq 0} \left\| \frac{B}{||B||^{(1)}} \right\|^{(2)} \\
&= ||A||^{(1)} \cdot \sup_{C:\|^{(1)}=1} \|C\|^{(2)} \\
&= ||A||^{(1)} \cdot \max_{C:\|C\|^{(1)}=1} \|C\|^{(2)},
\end{aligned}$$

where $\max_{C:\|C\|^{(1)}=1} \|C\|^{(2)} < \infty$ because the set $\left\{C \in M_N(\mathbb{C}) : \|C\|^{(1)} = 1\right\}$ is a compact set and the map $C \mapsto \|C\|^{(2)}$ is a continuous function. Thus, one can take the constant

$$c_2 = \max_{C:\|C\|^{(1)}=1} \|C\|^{(2)}.$$

Likewise, the constant c_1 can be chosen as $\max_{C:\|C\|^{(2)}=1} \|C\|^{(1)}$ □

Remark. *If $c_1||A||^{(1)} \leq ||A||^{(2)} \leq c_2||A||^{(1)}$ $\forall A \in M_N(\mathbb{C})$ then a sequence $A_n \in M_N(\mathbb{C})$ converges to the zero matrix under $||\cdot||^{(1)}$ if and only if $A_n \to 0$ under $||\cdot||^{(2)}$.*

In statistics, when given random variables X, Y with finite mean and variance, we can say $X, Y : \Omega \to \mathbb{R}$(or $\mathbb{C}$) are Borel measurable in probability space $(\Omega, \mathscr{F}, Pr)$.

Assume the expectations of X and Y, denoted by $E[X]$ and $E[Y]$ (abbreviated as EX

and EY for convenience) and variances of X and Y denoted by $Var(X)$ and $Var(Y)$ converge as Lebesgue integrals

$$EX = \int_{\Omega} X dPr,$$

$$EY = \int_{\Omega} Y dPr$$

$$Var(X) = E[(X - EX)^2] = \int_{\omega \in \Omega} (X(\omega) - EX)^2 dPr(\omega),$$

$$Var(Y) = E[(Y - EY)^2] = \int_{\omega \in \Omega} (Y(\omega) - EY)^2 dPr(\omega)$$

Definition (covariance). The covariance between random variables X and Y, denoted by $Cov(X, Y)$, is defined as:

$$Cov(X, Y) = E[(X - EX)(Y - EY)] \text{ for } X, Y : \Omega \to \mathbb{R}$$

or

$$Cov(X, Y) = E[(X - EX)\overline{(Y - EY)}] \text{ for } X, Y : \Omega \to \mathbb{C}$$

Definition (Covariance Matrix). X is a random vector on $\mathbb{R}^p$ if and only if X is a Borel measurable function on $\mathbb{R}$.

We can write

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

where each $x_i : \Omega \to \mathbb{R}$ is a real-valued random variable.

The covariance matrix C_X of random vector X is defined as

$$C_X = [c_{ij}]_{p \times p}$$

where $c_{ij} = Cov(x_i, x_j) = E[(x_i - Ex_i)(x_j - Ex_j)]$.

Remark. *1. By definition, $cov(x_i, x_j) = cov(x_j, x_i)$, meaning $c_{ij} = c_{ji}\ \forall i, j$, which implies that C is symmetric ($C = C^T$) in the real case, or self-adjoint in the complex case. Because*

of this, the spectral theorem applies to C.

2. *Let* $b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \in \mathbb{R}^p$ *and* $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ *be a random vector on* $\mathbb{R}^p$*, then*

$$b^T X = \begin{bmatrix} b_1 & b_2 & \dots & b_p \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \sum_{j=1}^{p} b_j X_j,$$

is a real-valued random variable.

Calculating the variance of $b^T X$*, we get:*

$$\begin{aligned} Var(b^T X) &= E\left[\left\{\sum_{j=1}^{p} b_j x_j - E\left(\sum_{j=1}^{p} b_j x_j\right)\right\}^2\right] \\ &= E\left[\left\{\sum_{j} b_j (x_j - Ex_j)\right\}\left\{\sum_{i} b_j (x_i - Ex_i)\right\}\right] \\ &= E\left[\sum_{i,j} b_i b_j (x_i - Ex_i)(x_j - Ex_j)\right] \\ &= \sum_{i,j} b_i Cov(x_i, x_j) b_j \\ &= \begin{bmatrix} b_1 & b_2 & \dots & b_p \end{bmatrix} C_X \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \\ &= b^T C_X b, \end{aligned}$$

which is the quadratic form associated with the matrix C_X*.*

By our assumptions earlier,

$$Var(b^T X) = \int_{\Omega} (b^T X - E b^T X)^2 dPr \geq 0$$

This shows that $b^T X b \geq 0 \forall\, b \in \mathbb{R}^p$, meaning the covariance matrix is positive semidefinite. If $b^T C_X b = 0$ for some $b \in \mathbb{R}^p$, then $Var(b^T X) = 0$ meaning $b^T X \sum_{j=1}^p b_j x_j = 0$ almost surely (i.e. there exists a set in Ω where $E \subset \Omega, Pr(E) = 0$ and $\sum_{j=1}^p b_j x_j(\omega) = 0 \quad \forall \omega \in \Omega \setminus E$). Meaning random vector X is "degenerate" (its distribution is supported on the hyper plane $\sum_{j=1}^p b_j t_j = 0$ if one parameterizes $\mathbb{R}^p = \{(t_1, t_2, \ldots, t_p) : t_1, \ldots, t_p \in \mathbb{R}\}$).

3. All eigenvalues of C are positive $\lambda_1 < \lambda_2 < \cdots < \lambda_p$, with eigenvectors $v_1, v_2, \ldots, v_p$ forming an orthonormal basis on $\mathbb{R}^p$. Any unit vector $x \in \mathbb{R}^p, x = \sum_{j=1}^p \alpha_j v_j$, with the operator norm of C, denoted by $||\cdot||$, we have:

$$\begin{aligned} ||Cx|| = ||\sum \alpha_j C v_j|| &\leq \lambda_1|\alpha_1| + \lambda_2|\alpha_2| + \cdots + \lambda_p|\alpha_p| \\ &< \lambda_p|\alpha_1| + \lambda_p|\alpha_2| + \cdots + \lambda_p|\alpha_p| \\ &= \lambda_p \sum |\alpha_j| \\ &\leq \lambda_p ||x||_{\mathbb{R}^p}, \end{aligned}$$

which implies that $||C|| \leq \lambda_p$, the maximum eigenvalue.
Conversely, $||Cv_p|| = \lambda_p$ implies that $||C|| = \lambda_p$.

We continue with an example in big data analysis. Let $p \approx \infty$. We may also say p is large.

Definition (Gaussian/Normal Density). Given $C \in M_p(\mathbb{R})$, where C is symmetric ($C^T = C$) and positive definite, $b^T C b > 0 \,\forall\, b \neq 0$ and $\mu \in \mathbb{R}^p$. Define the Gaussian density function for all $t = (t_1, t_2, \ldots, t_p) \in \mathbb{R}^p$

$$0 < f_{\mu,C}(t) = \frac{1}{(2\pi)^{p/2}} \frac{1}{\sqrt{detC}} ex\left(\frac{-1}{2}(t-\mu)C^{-1}(t-\mu)\right)$$

In the case that $\mu = 0 \in \mathbb{R}^p$ and C is the identity matrix $I_p \in \mathbb{R}^p$, the function $f_{0,C}$ is called the standard Gaussian density function.

Properties. 1. $f_{\mu,C}$ is integrable over $\mathbb{R}^p$ with respect to the Lebesgue measure $dt_1 dt_2 \ldots dt_p$ on $\mathbb{R}^p$ and

$$\int_{\mathbb{R}^p} f_\mu dt_1 dt_2 \ldots dt_p = 1$$

2. The Borel probability measure $\gamma_{\mu,C}$ on $\mathbb{R}^p$ defined by

$$\gamma_{\mu,C}(E) = \int_E f_{\mu,C} dt_1 dt_2 \dots dt_p, \forall \text{ Borel measurable } E \subset \mathbb{R}^p$$

is called the Gaussian (normal) distribution on $\mathbb{R}^p$.

3. In the case that $p = 1$, for $\mu \in \mathbb{R}, C > 0$,

$$\gamma_{\mu,C}(E) = \int_E \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{C}} \exp\left(\frac{(t-\mu)^2}{2C}\right) dt, \forall \text{ Borel measurable } E \subset \mathbb{R}.$$

Definition (Standard Gaussian/Normal Distribution). A random vector

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} : (\Omega, Pr) \to \mathbb{R}^p$$

is said to be a Gaussian random vector (said to have Gaussian distribution with mean μ and covariance C) if

$$Pr(\{\omega \in \Omega : x(\omega) \in E\}) = \gamma_{\mu,C}(E) \, \forall \text{ Borel measurable } E \subset \mathbb{R}^p.$$

Remark. *If a random vector* $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ *has standard Gaussian distribution, then* $x_1, x_2, \dots, x_p$ *are independently, identically distributed (i.i.d.) with* $\gamma_{0,1}$ *as the common distribution on* $\mathbb{R}$. *That is,* $\forall\, 1 \leq i \leq p$,

$$Pr(\{\omega \in \Omega : x_i(\omega) \in E\}) = \gamma_{0,1}(E) = \int_E \frac{1}{\sqrt{2\pi}} e^{t^2/2} dt$$

holds for all Borel sets $E \subset \mathbb{R}$.

1.2 The Large Dimension Paradox: The loss of matrix norm equivalence

In a statistical problem, people believe the underlying distribution is Gaussian $\gamma_{0,C}$ with covariance $C \in M_P(\mathbb{R})$. It would be of some interest to estimate C. We will do that by "collecting data" through using realizations of $X(\omega)$ for some $\omega \in \Omega$.

Theorem 1.2.1 (Strong Law of Large Numbers (SLLN)). (See Thm 22.1 in [3]) Suppose that $y_1, y_2, \ldots$ are i.i.d. real-valued random variables on a probability space Ω, and assume that they have finite mean $E[y_1]$. Then $\frac{1}{n}\sum y_i(\omega) \xrightarrow[n\to\infty]{} E[y_1]$ holds for almost all $\omega \in \Omega$.

To estimate C, we take $x_1, x_2, \ldots, x_n$ i.i.d. random vectors with a common distribution $\gamma_{0,C}$, that is, $Pr(x_j \in E) = \gamma_{0,C}(E)$ for all Borel sets $E \in \mathbb{R}^P$. Then we form a sample covariance matrix

$$\hat{C} = \frac{1}{n} X X^T = \frac{1}{n} \sum_{j=1}^{n} x_j x_j^T.$$

Here we let $X = [X_1 X_2 \ldots X_n] = [x_{ij}]$ where $X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{pj} \end{bmatrix}$.

Then,

$$\begin{aligned} \frac{1}{n} X X^T &= \frac{1}{n} [x_{ij}][x_{ji}] \\ &= \frac{1}{n} [\hat{C}_{ij}], \end{aligned}$$

where

$$\hat{C}_{ij} = \sum_{k=1}^{n} x_{ik} x_{jk}.$$

Example 1.2.1. Calculating $\hat{C_{11}}$, we get:

$$\begin{aligned}\hat{C}_{11} &= x_{11}x_{11} + x_{12}x_{12} + x_{13}x_{13} + \cdots + x_{1n}x_{1n} \\ &= x_{11}^2 + x_{12}^2 + x_{13}^2 + \cdots + x_{1n}^2\end{aligned}$$

which is a sum of i.i.d. real-valued random variables.

Given Borel set $E \subset \mathbb{R}$

$$\begin{aligned}Pr(x_{11} \in E) &= E[\phi(x_1)] \text{ where } \phi = I_E \times I_{\mathbb{R}} \times I_{\mathbb{R}} \times \cdots \times I_{\mathbb{R}} \\ &= E[\phi(x_j)].\end{aligned}$$

Example 1.2.2. Calculating $\hat{C_{21}}$, we get:

$$\hat{C}_{21} = x_{21}x_{11} + x_{22}x_{12} + \cdots + x_{2n}x_{1n},$$

which is a sum of i.i.d. random variables by our assumption.

In general, SLLN can be applied when we fix the variable P.

$$\begin{aligned}\frac{1}{n}\hat{C}_{ij}(\omega) &\xrightarrow[n\to\infty]{} E[x_{ik}x_{jk}] \\ &\overset{i.i.d.}{=} E[x_{i1}x_{j1}] \\ &= Cov(x_{i1}, x_{j1}) \\ &\overset{def}{=} C_{ij}\end{aligned}$$

for almost all $\omega \in \Omega$.

In conclusion, we have the almost-sure entry-wise convergence of $\hat{C} \xrightarrow[n\to\infty]{} C$ when p is fixed. In particular, any matrix norm of $\hat{C} - C$ tends to zero almost surely as $n \to \infty$, including the operator norm $||\cdot||$. In other words, $||\hat{C} - C|| \xrightarrow[n\to\infty]{} 0$ holds for almost all $\omega \in \Omega$.

In big data analysis, p is as large as n (the number of random trials to estimate C) [4]. The main issue arises when $n, p \to \infty$ with $\frac{p}{n} \to c \in (0, \infty)$, the operator norm approximation fails. For the remainder of this section, we assume $c > 1$.

Theorem 1.2.2. (Corollary 1, Sec. 1.5 [1]) Let A be a Banach algebra with unit, then

the set of all invertible elements in A forms an open set under norm topology. Precisely, if $x_0 \in A$ is invertible, then there exists $\epsilon_0 > 0$ so that every x satisfying $||x - x_0|| < \epsilon_0$ is invertible.

We refer to [1] for a proof of the above theorem.

Proof. Note: $M_p(\mathbb{R})$ is a Banach algebra with the unit I_p under the operator norm $||\cdot||$ and the usual matrix multiplication.

Suppose, in order to derive a contradiction, that

$$\lim_{\substack{n,p\to\infty \\ p/n\to c}} ||\hat{C}(\omega) - C|| = 0,$$

then by Theorem 1.2.2, the invertibility of C implies that $\hat{C}(\omega)$ must also be invertible for large n and p.

Note the fact that X_j is Gaussian implies that X_j has absolutely continuous distribution, meaning $Pr(X_j = 0) = 0$. Thus,

$$X_j(\omega) = \begin{bmatrix} x_{1j}(\omega) \\ x_{2j}(\omega) \\ \vdots \\ x_{pj}(\omega) \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^p \text{ for almost all } \omega \in \Omega.$$

Say, some $x_{ij}(\omega) \neq 0$ for some $i \in [p] = \{1, 2, \ldots, p\}$, then

$$X_j(\omega)X_j(\omega)^T = \begin{bmatrix} x_{1j}(\omega)x_j(\omega)^T \\ x_{2j}(\omega)x_j(\omega)^T \\ \vdots \\ x_{ij}(\omega)x_j(\omega)^T \\ \vdots \\ x_{pj}(\omega)x_j(\omega)^T \end{bmatrix}$$

has rank 1.

Then

$$\hat{C}(\omega) = \frac{1}{n}\sum_{j=1}^{n} x_j(\omega)x_j(\omega)^T$$

has at most rank n. However, since $\frac{p}{n} \to c > 1$, we know $p > n$ for sufficiently large p, n, by the Rank Nullity Theorem, the rank of $\hat{C}(\omega)$ is at least $p - n \geq 1$, meaning $\hat{C}(\omega)$ cannot be invertible for large p and n, creating a contradiction. Thus, $\hat{C}$ is not a good estimator for C and the operator norm approximation fails. □

Moreover, as stated in [4], due to the concentration inequalities of Gaussian entries, the entry-wise approximation $\hat{C_{ij}}(\omega) \to C_{ij}$ holds as $n \to \infty$ uniformly in p so any matrix norm approximation still holds in this case. The failure of the operator norm approximation implies the matrix norm is not equivalent to the operator norm in this case. In view of this, it is natural to ask:

Is there any controllable asymptotic behavior for the sample covariance matrix $\hat{C}$?

We will continue this discussion in the next chapter with the introduction of the Marchenko-Pastur Law.

2 Marchenko-Pastur Law for sample covariance matrix of Gaussian entries

Let X denote a rectangular $p \times n$ random matrix of standard complex Gaussian entries. The Marchenko-Pastur Law for the sample covariance matrix $\hat{C} = \frac{1}{p}XX^*$ (where X^* denotes the conjugate transpose of X) is the following limit theorem concerning the arranged eigenvalues of $\hat{C}$:

Theorem 2.0.1 (Marchenko-Pastur Law). (From [6]) Denote by $\lambda_1, \lambda_2, \ldots, \lambda_p$ the random eigenvalues of $\hat{C}$, then as $n, p \to \infty$ with $\frac{p}{n} \to c \in (0, \infty)$, the averaged eigenvalue distribution denoted by

$$\mu_{\hat{C}} = E\left[\frac{1}{p}\sum_{j=1}^{p}\delta_{\lambda_j}\right]$$

converges weakly to a deterministic, absolutely continuous Borel probability measure γ_c, defined by the density

$$\left(1 - \frac{1}{c}\right)^+ \delta_0 + \frac{1}{2\pi c X}\sqrt{(x-1)(b-x)},$$

where $a = (1 - \sqrt{c})^2$, $b = (1 + \sqrt{c})^2$ and $(x)^+ = \max\{X, 0\}$.

We consider the sample covariance matrix of more general complex Gaussian entries, the result of the real Gaussian entries can be shown in the same way. In Chapter 3, we will discuss the Marchenko-Pastur Law for general entries, not necessarily Gaussian, as well as the stronger, almost-sure convergence of the eigenvalue distribution.

2.1 Preliminaries

2.1.1 Non-crossing Partitions and Permutations

Definition (Non-crossing partitions). (From [7]) Let π be a partition of $[n]$. If we can find $i < j < k < l$ such that i and k are in one block V and j and l are in another block W of π, we say V and W *cross*. If no pair of blocks of π cross, then we say π is *non-crossing*. We denote the set of non-crossing partitions of $[n]$ by $NC(n)$. The set of non-crossing pairings of $[n]$ is denoted $NC_2(n)$.

Definition (Permutations). (From [8]). Denote by S_n the symmetric group of permutations of $\{1, \ldots, n\}$

Proposition 2.1.2. (Proposition 23.11 of [8]) For $\sigma \in S_n, |\sigma| + \#(\sigma) = n$.

Proposition 2.1.3. (Proposition 23.22 of [8] and [2]) If γ denotes the permutation $(1\,2\,3 \ldots k)$ then $\#(\gamma\sigma) + \#(\sigma) = k - 1 \iff \sigma \in NC(k)$

2.1.2 Wick's Formula

Recall the fact that Z is a standard complex Gaussian random variable if and only if

$$Z = \frac{1}{\sqrt{2}} X + iY,$$

where X, Y are independent real standard Gaussian random variables, so

$$E[Z] = 0$$

and

$$E[Z\bar{Z}] = \frac{E[X^2]E[Y^2]}{2} = 1.$$

The joint distribution of X, Y is the probability measure $\gamma_{0,C}$ where the covariance matrix $C = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$.

For $m, n \in \mathbb{N}$

$$\begin{aligned} E[Z^m \bar{Z}^n] &= \int_{\mathbb{R}^2} (s+it)^m (s-it) \frac{1}{\pi} \exp\left(\frac{-2(s^2+t^2)}{2}\right) ds dt \\ &= \frac{1}{\pi} \int_0^{2\pi} e^{i(m-n)\theta} d\theta \\ &= \int_0^{\infty} r^{m+n+1} e^{-r^2} dr \\ &= \begin{cases} 0 \text{ if } m \neq n. \\ m! \text{ if } m = n. \end{cases} \end{aligned}$$

For mixed moments of general Gaussians, we have

Theorem 2.1.1 (Wick's Formula). (Corollary 2, Section 1.5 of [7]) Suppose $(X_1, \ldots, X_n)$ is a complex Gaussian random vector then

$$E(X_{i_1}^{\epsilon_1} \ldots X_{i_k}^{\epsilon_k}) = \sum_{\pi \in \mathcal{P}_2(k)} E_\pi(X_{i_1}^{\epsilon_1}, \ldots, X_{i_k}^{\epsilon_k})$$

for all $i_1, \ldots, i_k \in [n]$ and all $\epsilon_1, \ldots \epsilon_k \in \{0, 1\}$; where we have used the notation $X_i^{(0)} := X_i$ and $X_i^{(1)} := \overline{X_i}$.

Applying Wick's formula to independent standard complex Gaussians, we have the following results (see [7]):

Proposition 2.1.4. (Exercise 7 of [7])
If $Z_1, Z_2, \ldots, Z_s$ are independent, standard complex Gaussian random variables with $E[Z_j] = 0$ and $E[|Z_j|^2] = 1$, then

$$E[Z_{i1} Z_{i2} \ldots Z_{im} \bar{Z}_{j1} \bar{Z}_{j2} \ldots \bar{Z}_{jn}] = \begin{cases} 0 \text{ if } m \neq n \\ \#\{\sigma \in S_n : i = j \circ \sigma\} \text{ if } m = n \end{cases}$$

Proof. (exercise 6 on Part 5 of [7])
The balance condition in proposition 2.1.4 shows that each Z_i has to be paired with some $\bar{Z}_j$. In other words if $m \neq n$, then the expression equals 0.

When $m = n$, Wick's formula shows:

$$\begin{aligned} E[*] &= \sum_{\substack{\pi \in \mathcal{P}_2(m+n) \\ \pi \text{ pairs } Z_i with \bar{Z}_j}} \prod_{(r,s)\in\pi} E[Z_{ir}\bar{Z_{js}}] \\ &= \sum_{\substack{\pi \in \mathcal{P}_2(m+n) \\ \pi \text{ pairs } Z_i with \bar{Z}_j}} 1 \\ &= \#\{\text{all such } \pi\} \end{aligned}$$

Every such π determines a unique permutation $\sigma \in S_n$ as follows:

$$\begin{aligned} Z_{i_1} &\longrightarrow \overline{Z_{j_{\sigma(1)}}} \\ Z_{i_2} &\longrightarrow \overline{Z_{j_{\sigma(2)}}} \\ &\vdots \\ Z_{i_n} &\longrightarrow \overline{Z_{j_{\sigma(n)}}} \end{aligned}$$

Thus σ satisfies $i = j \circ \sigma$ on $[n] = \{1, 2, \ldots n\}$.

Conversely, if $\sigma \in S_n$ satisfies $i = j \circ \sigma$, then we write out π as follows:

$$Z_{i_1} Z_{i_2} \ldots Z_{i_k} \ldots Z_{i_n} \overline{Z_{j_1} Z_{j_2}} \ldots \overline{Z_{j_{\sigma(k)}}} \ldots \overline{Z_{j_n}}$$

Thus, $E[*] = \#\{\text{all such } \pi\} = \#\{\sigma \in S_n : i = j \circ \sigma\}$ □

2.1.3 Free Cumulants and Cauchy Transform

Definition (Free cumulants and the moment-cumulant formula). (Definition 8, Section 2.2 of [7]). Let $(\mathcal{A}, \phi)$ be a non-commutative probability space. The corresponding free cumulants denoted by $\kappa_n : \mathcal{A}^n \to \mathbb{C} (n \geq 1)$ are defined inductively in terms of moments by the moment-cumulant formula:

$$\phi(a_1 \ldots a_n) = \sum_{\pi \in NC(n)} \kappa_\pi(a_1, \ldots, a_n)$$

where, by definition, if $\pi = \{V_1, \ldots, V_r\}$, then

$$\kappa_\pi(a_1 \ldots a_n) = \prod_{\substack{V \in \pi \\ V = (i_1 \ldots i_l)}} \kappa_l(a_{i_1}, \ldots a_{i_l})$$

Definition (The case when $a_1 = a_2 = \cdots = a_n = a$). (2.16 of [7])

$$\phi(a^n) = \sum_{\pi \in NC(n)} \kappa_\pi^a = \prod_{\pi = V_1 \cup \cdots \cup V_r} \kappa_{|V_j|}(a, a, \ldots, a)$$

Remark. *If a, b are free, then the free cumulant is additive:*

$$\kappa_n^{a,b} = \kappa_n^a + \kappa_n^b \; \forall n.$$

Proposition 2.1.5. (Section 17 section 2.4 of [7]) The relation between the moment series $M(z)$ and the cumulant series $C(z)$ of a random variable is given by

$$M(z) = C(zM(z)).$$

Assume $\phi(a^n) = \int_{\mathbb{R}} t^n d\mu(t), \forall n = 1, 2, \ldots,$ where μ is a probability measure on $\mathbb{R}$, uniquely determined by its moments, then

$$G_a(z) = \int_{\mathbb{R}} \frac{1}{z - t} d\mu(t), z = x + iy, y > 0$$

is called the Cauchy transform of μ.

Properties. Let v be a probability measure on $\mathbb{R}$ with Cauchy transform G.

1. (Lemma 3, Chapter 3 of [7])

$$\lim_{y \to \infty} iyG(iy) = 1$$

and

$$\sup_{y>0, x \in \mathbb{R}} y|G(x + iy)| = 1.$$

2. (Theorem 6, Chapter 3 of [7]) For $a < b$ we have

$$-\lim_{y\to 0^+}\frac{1}{\pi}\int_a^b \Im(G(x+iy))dx = v((a,b)) + \frac{1}{2}v(\{a,b\}).$$

If v_1 and v_2 are probability measures with $G_{v_1} = G_{v_2}, then v_1 = v_2$.

3. (Proposition 8, Chapter 3 of [7]) For all $a \in \mathbb{R}$ we have

$$\lim_{z\to a}(z-a)G(z) = v(\{a\}).$$

2.1.4 The Method of Moments

Theorem 2.1.2 (Theorem 30.2 of [3]). Let $\mu, \mu_n, (n \in \mathbb{N})$ be probability measures on $\mathbb{R}$. Assume that μ is uniquely determined by its moments, and

$$\lim_{n\to\infty}\int_{\mathbb{R}} t^k d\mu_n(t) = \int_{\mathbb{R}} t^k d\mu(t), \forall k = 1, 2, 3, \ldots$$

then μ_n converges weakly to μ as $n \to \infty$.

2.2 Proof of the Marchenko-Pastur Law

2.2.1 Convergence in Moments

Recall $X = [Z_{ij}]_{p\times n}$ is a rectangular random matrix with independent standard complex Gaussians Z_{ij}. First, we will investigate $\frac{1}{p}XX^* = Y$. Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the random eigenvalues of Y and set $tr(A) = \frac{1}{p}\operatorname{Tr}(A)\, \forall A \in M_p(\mathbb{C})$. The average eigenvalue distribution is

$$\mu_Y = E\left[\frac{1}{p}\sum_{j=1}^{p}\delta_{\lambda_j}\right]$$

By the spectral theorem, we have the k-th moment

$$\begin{aligned}\int_{\mathbb{R}} t^k d\mu_Y(t) &= E[tr(Y^k)]] \\ &= \frac{1}{p^{k+1}} E[\mathrm{Tr}(\underbrace{XX^*XX^*\dots XX^*}_{\text{2k terms}})] \\ &= p^{-(k+1)} \sum_{\substack{i:[k]\to[p] \\ j:[k]\to[n]}} \underbrace{E[Z_{i_1j_1}\overline{Z_{i_2j_1}}Z_{i_2j_2}\overline{Z_{i_3j_2}}\dots Z_{i_kj_k}\overline{Z_{i_1j_k}}]}_{(*)}\end{aligned}$$

Let γ denote the one-cycle permutation $(1\,2\,3\dots k)$. Then

$$\begin{aligned}(*) &= E[Z_{i_1j_1}Z_{i_2j_2}\dots Z_{i_kj_k}\overline{Z_{i_2j_1}Z_{i_3j_2}}\dots\overline{Z_{i_1j_k}}] \\ &= card\{\sigma\in S_k:\ \text{each } i_lj_l \text{ pairs with a unique } i_{\sigma(l)+1}j_{\sigma(l)} = i_{\gamma(\sigma(l))}j_{\sigma(l)} \text{ via } \sigma\}\end{aligned}$$

In other words, we have to count all such $\sigma\in S_k$ through the conditions: $i = i\circ(\gamma\sigma), j = j\circ\sigma$ on the set $[k] = \{1,2,\dots,k\}$. We conclude from the conditions that i must remain a constant on each cycle of $\gamma\sigma$ and so does j on each cycle of σ.

Since there are $p^{\#(\gamma\sigma)}$ many i's and $n^{\#(\sigma)}$ many j's, we obtain

$$\begin{aligned}m_k\int_{\mathbb{R}} t^k d\mu_Y(t) &= p^{-(k+1)}\sum_{\sigma\in S_k} p^{\#(\gamma\sigma)}\cdot n^{\#(\sigma)} \\ &= \sum_{\sigma\in S_k} p^{\#(\gamma\sigma)-k-1+\#(\sigma)}\cdot\left(\frac{n}{p}\right)^{\#(\sigma)}\end{aligned}$$

Now, recall the assumptions: $n,p\to\infty$, $\dfrac{p}{n}\to c>0 \Rightarrow \dfrac{n}{p}\to\dfrac{1}{c} = d > 0$. Note that it is easy to see from proposition 2.1.3 that $\#(\gamma\sigma)+\#(\sigma)\le k+1$ and the equality holds if and only if $\sigma\in NC(k)$. So in the limit $n,p\to\infty$ with $\dfrac{n}{p}\to d>0$, we conclude the convergence of the k-th moment $m_k\to\sum_{\sigma\in NC(k)} d^{\#(\sigma)}$.

In the next subsection, we will identify this limit as the k-th moment of a probability measure on $\mathbb{R}$.

2.2.2 Limit Distribution

In the previous section, we proved that k-th moment converges $m_k\to\sum_{\sigma\in NC(k)} d^{\#(\sigma)}$. If we take a non-commutative random variable x with constant free cumulants $\kappa_n^x = d$ for $n = 1,2,\dots$, then the moment-cumulant formula shows that $\phi(x^k) = \sum_{\sigma\in NC(k)} d^{\#(\sigma)}$. Thus,

we have the R-transform of x given by $R_x(z) = d + dz + dz^2 + \cdots = \dfrac{d}{1-z} = G_x^{-1}(z) - \dfrac{1}{z}$, which further implies that

$$G_x^{-1}(z) = \frac{d}{1-z} + \frac{1}{z} = \frac{dz + 1 - z}{z - z^2}.$$

It follows that the Cauchy transform is either

$$G_x(z) = \frac{z - (d-1) + \sqrt{[z-(d-1)]^2 - 4z}}{2z}$$

or

$$G_x(z) = \frac{z - (d-1) - \sqrt{[z-(d-1)]^2 - 4z}}{2z}$$

where the branch of the square root is chosen as follows:

$$\sqrt{z} = \sqrt{r}e^{i\theta/2} \text{ for } z = re^{i\theta}, \theta \in [0, 2\pi).$$

By property (1) on the Cauchy transform in subsection 2.1.3, the correct form of G_x is given by

$$G_x(z) = \frac{z - (d-1) - \sqrt{[z-(d-1)]^2 - 4z}}{2z}$$

for z in the upper half-plane.

Note that

$$\begin{aligned}
[z-(d-1)]^2 - 4z &= z^2 - 2(d-1)z + (d-1)^2 - 4z \\
&= z^2 - 2(d+1)z + (d+1)^2 - 4d \\
&= [z-(d+1)]^2 - 4d.
\end{aligned}$$

Therefore, we have

$$G_x(z) = \frac{z - (d-1) - \sqrt{(z-a)(z-b)}}{z}$$

provided that $a = (d+1) - 2\sqrt{d} = (\sqrt{d}-1)^2$ and $b = (d+1) + 2\sqrt{d} = (\sqrt{d}+1)^2$.

Note that G_x is precisely the Cauchy transform of the free Poisson distribution with parameter d.

Recall that $Y = \dfrac{1}{p}XX^*$ and

$$E[tr(Y^k)] \xrightarrow[\substack{n,p\to\infty \\ p/n\to c>0}]{} \phi(x^k) = \int_{\mathbb{R}} t^k d\mu_d(t),$$

then the k-th moment of $\hat{C} = \frac{1}{n}XX^*$:

$$\begin{aligned} E[tr(\hat{C}^k)] &= E[tr((\frac{p}{n}Y)^k)] \\ &= \left(\frac{p}{n}\right)^k E[tr(Y^k)] \xrightarrow[\substack{n,p\to\infty \\ p/n\to c>0}]{} c^k\phi(x^k) \\ &= c^k \int_{\mathbb{R}} t^k d\mu_d(t) \\ &= \int_{\mathbb{R}} (ct)^k d\mu_d(t) \\ &= \int_{\mathbb{R}} t^k d\nu_c(t) \end{aligned}$$

where ν_c denotes the pushforward measure of the free Poisson law μ_d via the transformation $T(t) = ct, t \in \mathbb{R}$, and ν_c is precisely the distribution of the dilation cx.

Note that the Cauchy transform of cx satisfies

$$G_{cx}(z) = \frac{1}{c}G_x\left(\frac{z}{c}\right), \forall \Im z > 0.$$

Thus by properties (2), (3) on Cauchy transform, we have:

1. $\nu_c(\{0\}) > 0 \iff c > 1$, meaning $\nu_c(\{0\}) = 1 - \frac{1}{c}$
2. The support of ν_c is the interval $[A, B]$ where $A = ca = (1-\sqrt{c})^2$, $B = cb = (1+\sqrt{c})^2$.
3. Denote by m the Lebesgue measure on $\mathbb{R}$. The density ν_c is given by the dilation

$$\frac{d\nu_c}{dm}(t) = \frac{1}{c}\frac{d\mu_d}{m}\left(\frac{t}{c}\right)$$

for m-almost all $t \in \mathbb{R}$.

In summary, we have

$$d\nu_c(t) = \left(1 - \frac{1}{c}\right)^+ \delta_0 + \frac{\sqrt{(t-A)(B-t)}}{2\pi ct} dt \text{ for } t \in [A, B]$$

.

2.2.3 Convergence in Distribution

Denote by μ_n the averaged eigenvalue distribution of $\hat{C} = \frac{1}{n}XX^*$. We have shown that μ_n converges in moments to ν_c as $n, p \to \infty$ with $\frac{p}{n} \to c$.

We will now show that μ_n converges weakly to ν_c. In order to apply the method of moments, we need to show that ν_c is uniquely determined by its moments. Quoting from Theorem 30.1 of [3] and Lemma 3.2.6 of [5] as follows:

Theorem 2.2.1 (Theorem 30.1 of [3] and Lemma 3.2.6 of [5]). If μ is a Borel Probability measure on $\mathbb{R}$ such that $\int_{\mathbb{R}} |t|^k d\mu(t) < \infty \forall k = 1, 2, \ldots$ and the power series $\sum_{k=1}^{\infty} \frac{\alpha_k}{k!} r^k$ has the radius of convergence $R = \dfrac{1}{limsup \sqrt[k]{\dfrac{|\alpha_k|}{k!}}} > 0$, where $\alpha_k \int_{\mathbb{R}} t^k d\mu(t)$, then μ is uniquely determined by its moments.

Applying this to ν_c, we have

$$\sqrt[k]{k!} \geq 1, k = 1, 2, \ldots$$

and

$$\sqrt[k]{\left| \int_{\mathbb{R}} t^k d\nu_c(t) \right|} = \sqrt[k]{\int_A^B t^k d\nu_c(t)} \leq \sqrt[k]{B^k} = B$$

and hence,

$$\limsup_{k \to \infty} \sqrt[k]{\frac{|\alpha_k|}{k!}} \leq B,$$

thus the radius of convergence is

$$R = \frac{1}{\limsup\limits_{k}} \geq B > 0.$$

So, ν_c is uniquely determined by its moments and theorem implies $\mu_n \Rightarrow \nu_c$.

3 Further discussion

First, concerning the stronger, almost-sure convergence of the Marchenko-Pastur Law, Marchenko and Pastur actually show in the original result from [6] that the law also holds for entries with zero mean and finite variance (with minor moment conditions), not necessarily Gaussian entries.

Secondly, the law in fact holds almost surely for the "un-averaged" eigenvalue distribution $\frac{1}{p}\sum_{j=1}^{p}\delta_{\lambda_j}$ with the same limit law ν_C.
For these general results, we quote Theorem 2.4 from the book [4] as follows:

Theorem 3.0.1 ((Theorem 2.4 of [4]). Let $X \in \mathbb{R}^{p\times n}$ with i.i.d. columns x_i such that x_i has independent entries with zero mean, unit variance, and some light tail condition and denote the resolvent of $\frac{1}{n}XX^T$ as

$$Q(z) = \left(\frac{1}{n}XX^T - zI_p\right)^{-1}.$$

Then as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$,

$$Q(z) \leftrightarrow \overline{Q}(z),\ \overline{Q}(z) = m(z)I_p$$

with $(z, m(z))$ the unique solution in $Z(\mathbb{C} \setminus [(1-\sqrt{c})^2, (1+\sqrt{c})^2])$ of

$$zcm^2(z) - (1-c-z)m(z) + 1 = 0.$$

The function m(z) is the Stieltjes transform of the probability measure μ given explicitly by

$$\mu(dx) = (1-c^{-1})^+\delta_0(x) + \frac{1}{2\pi cx}\sqrt{(x-E_-)^+(E_+-x)^+}dx$$

where $E_\pm = (1\pm\sqrt{c})^2$ and $(x)^+ = \max(0, x)$, and is known as the Marchenko Pastur distribution. In particular, with probability one, the empirical spectral measure $\mu_{\frac{1}{n}XX^T}$ converges weakly to μ.

One can even go beyond measures with finite variance, see [4] and the references therein.

References

[1] William Arveson. *A short course on spectral theory*, volume 209 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2002.

[2] Philippe Biane. Some properties of crossings and partitions. *Discrete Math.*, 175(1-3):41–53, 1997.

[3] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[4] Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022.

[5] Robert E. Greene and Steven G. Krantz. *Function theory of one complex variable*. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, 1997. A Wiley-Interscience Publication.

[6] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):507–536, 1967.

[7] James A. Mingo and Roland Speicher. *Free probability and random matrices*, volume 35 of *Fields Institute Monographs*. Springer, New York; Fields Institute for Research in Mathematical Sciences, Toronto, ON, 2017.

[8] Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 335 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2006.