Random Matrix Theory and Its Applications in ML: Part 1 Random Matrix Theory

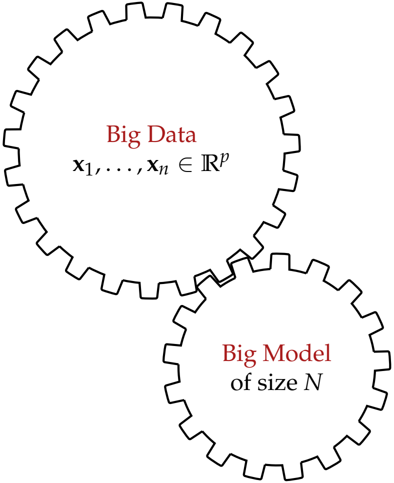
Short Course @ Jiangsu Normal University

Zhenyu Liao

School of Electronic Information and Communications
Huazhong University of Science and Technology

January 11, 2024

Outline



- **Big Data era**: exploit large n, p, N
- counterintuitive phenomena different from classical asymptotics statistics
- complete change of understanding of many methods in statistics, machine learning, signal processing, and wireless communications
- **Random Matrix Theory (RMT)** provides the tools!

Sample covariance matrix in the large n, p regime

- **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,
- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \in \mathbb{R}^{p \times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

almost surely as $n \to \infty$: optimal for $n \gg p$ (or, for p "small").

- In the regime $n \sim p$, conventional wisdom breaks down: for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has at least $p - n$ zero eigenvalues:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{ eigenvalue mismatch and not consistent!}$$

- due to $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\|\mathbf{A}\|_\infty \equiv \max_{ij} |\mathbf{A}_{ij}|$.

When is one in the random matrix regime? Almost always!

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: MP law

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

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

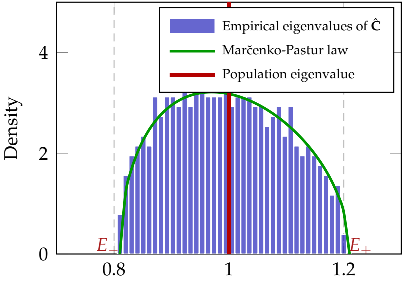


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[E_- = (1 - \sqrt{\mathbf{c}})^2, E_+ = (1 + \sqrt{\mathbf{c}})^2]$.
- for $\mathbf{n} = \mathbf{100p}$, on a range of $\pm 2\sqrt{\mathbf{c}} = \pm \mathbf{0.2}$ around the population eigenvalue $\mathbf{1}$.

Classical large-n asymptotic analysis mostly fails today

- large-n intuition, and many existing popular methods in biology, finance, signal processing, telecommunication, and machine learning, must **fail** even with $n = 100p$!
- **RMT** as a flexible and powerful tool to **understand** and **recreate** these methods
- in essence, "increasing complexity of the system models employed in above fields demand low complexity analysis"
- as motivating examples, how RMT can be applied to assess:
- **telecommunication**: code division multiple access (CDMA) technology
- **signal processing**: generalized likelihood ratio test (GLRT)
- **machine learning**: principle component analysis (PCA), and kernel spectral clustering

Application to telecom: performance analysis of CDMA via RMT

- **CDMA**: code division multiple access, key technology in 3G
- **Idea**: to increase max number of users, and dynamically balancing the quality of service to each terminal
- each user is allocated a (long) **spreading code** orthogonal to the other users' codes
- all users can simultaneously receive data while experiencing a limited amount of interference from concurrent communications, due to code orthogonality
- codes **not** fully orthogonal, more users, more interference and less quality of service
- **Question**: how to evaluate the **capacity** (max achievable transmission data rate) of CDMA network? (which clearly depends on pre-coding strategy)

Orthogonal CDMA versus TDMA

- for **orthogonal** CDMA, under some commonly used technical assumptions, capacity given by

$$C_{\rm orth}(\sigma^2) = \frac{1}{n} \log\det\left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{W}\mathbf{G}\mathbf{G}^{\sf H}\mathbf{W}^{\sf H} \right), \tag{1}$$

  with noise power σ^2, $\mathbf{W} \in \mathbb{C}^{n\times n}$ the **orthogonal** CDMA codes ($\mathbf{W}$ unitary), and $\mathbf{G} \equiv \mathrm{diag}\{g_i\}_{i=1}^n$ represents channel **gains** of the users.
- Note

$$C_{\rm orth}(\sigma^2) = \frac{1}{n} \log\det\left(\mathbf{I}_n + \frac{1}{\sigma^2} \mathbf{G}\mathbf{G}^{\sf H} \right) = \frac{1}{n} \sum_{i=1}^n \log\left(1 + \frac{|g_i|^2}{\sigma^2} \right) = C_{\rm TDMA}(\sigma^2), \tag{2}$$

  justifies the equivalence between TDMA (for 2G) and **orthogonal** CDMA rate performance.

Random versus orthogonal CDMA

- however, orthogonality can be **computationally demanding**: random CDMA with random i.i.d. codes,

$$C_{\mathrm{rand}}(\sigma^2) = \frac{1}{n} \log\det\left(\mathbf{I}_n + \frac{1}{\sigma^2}\mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}}\right), \tag{3}$$

  for $\mathbf{X} \in \mathbb{C}^{n \times n}$ the users' random codes.
- from RMT perspective, denote μ the **empirical spectral measure** of $\mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}}$, then $C_{\mathrm{rand}}(\sigma^2) = \int \log(1 + t/\sigma^2)\mu(dt)$: known as **linear spectral statistics** (LSS) of $\mathbf{X}\mathbf{G}\mathbf{G}^{\mathsf{H}}\mathbf{X}^{\mathsf{H}}$
- **Question**: C_{rand} as a function of gains $\mathbf{G}$ and (distribution of) codes $\mathbf{X}$?
- (first?) answered by Shami, Tse, and Verdú in [TV00; VS99];
- however capacity expressions not achievable in practice, due to complicated and nonlinear processing
- if only linear pre-coders and/or decoders are used, optimal solution:
  - frequency flat channels [TH99]: D.N.C. Tse and S.V. Hanly. "Linear multiuser receivers: effective interference, effective bandwidth and user capacity". In: *IEEE Transactions on Information Theory* 45.2 (1999), pp. 641–657
  - frequency selective channels [ET00]: J. Evans and D.N.C. Tse. "Large system performance of linear multiuser receivers in multipath fading channels". In: *IEEE Transactions on Information Theory* 46.6 (2000), pp. 2059–2078
  - reduced-rank LMMSE decoders [LTV04]: Linbo Li, Antonia M. Tulino, and Sergio Verdú. "Design of Reduced-Rank MMSE Multiuser Detectors Using Random Matrix Methods". In: *IEEE Transactions on Information Theory* 50.6 (2004), pp. 986–1008
  - etc.

Signal sensing using multi-dimensional sensor arrays

Motivation:

- Shannon: to achieve high rate of information transfer, increasing the transmission bandwidth is largely preferred over increasing the power
- high rate communications with finite power budget, need **frequency multiplexing**
- **cognitive radio**: to communicate not by exploiting the over-used frequency domain, or by exploiting the over-used space domain, but by exploiting so-called **spectrum holes**, jointly in time, space, and frequency

As such, a cognitive radio network (also called a *secondary network*)

- can help reuse the resources in a licensed (*first*) network
- but require constant **awareness** of the operations taking place in the licensed networks
- for example, via **signal sensing/detection**

Hypothesis testing in a signal-plus-noise model for cognitive radios

System model: let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ with i.i.d. columns $\mathbf{x}_i \in \mathbb{R}^p$ received by array of p sensors, signal decision as the following binary hypothesis test:

$$\mathbf{X} = \begin{cases} \sigma \mathbf{Z}, & \mathcal{H}_0 \\ \mathbf{a}\mathbf{s}^\mathsf{T} + \sigma \mathbf{Z}, & \mathcal{H}_1 \end{cases}$$

where $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p \times n}$, $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$ deterministic of unit norm $\|\mathbf{a}\| = 1$, signal $\mathbf{s} = [s_1, \ldots, s_n]^\mathsf{T} \in \mathbb{R}^n$ with s_i i.i.d. random, and $\sigma > 0$. Denote $c = p/n > 0$.

- observation of either zero-mean Gaussian **noise** $\sigma \mathbf{z}_i$ of power σ^2, or deterministic **information** vector $\mathbf{a}$ modulated by an added scalar (random) **signal** s_i (e.g., ± 1).
- If $\mathbf{a}$, σ, and statistics of s_i are known, the decision-optimal Neyman-Pearson () test:

$$\frac{\mathbb{P}(\mathbf{X} \mid \mathcal{H}_1)}{\mathbb{P}(\mathbf{X} \mid \mathcal{H}_0)} \overset{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \alpha \tag{4}$$

  for some $\alpha > 0$ controlling the Type I and II error rates.

Hypothesis testing via GLRT

However,

- in practice, we do not know σ, nor the information vector $\mathbf{a} \in \mathbb{R}^p$ (to be recovered)
- in the case of $\mathbf{a}$ fully unknown, one may resort to a **generalized likelihood ratio test** (GLRT) defined as

$$\frac{\sup_{\sigma,\mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathbf{a}, \mathcal{H}_1)}{\sup_{\sigma,\mathbf{a}} \mathbb{P}(\mathbf{X} \mid \sigma, \mathcal{H}_0)} \overset{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} \alpha.$$

- Gaussian noise and signal s_i, GLRT has an explicit expression as a monotonous increasing function of $\|\mathbf{X}\mathbf{X}^\mathsf{T}\| / \operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})$, test equivalent to, for some known f,

$$T_p \equiv \frac{\|\mathbf{X}\mathbf{X}^\mathsf{T}\|}{\operatorname{tr}(\mathbf{X}\mathbf{X}^\mathsf{T})} \overset{\mathcal{H}_1}{\underset{\mathcal{H}_0}{\gtrless}} f(\alpha).$$

- to evaluate the power of GLRT above, we need to assess the max and mean eigenvalues of SCM $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$

Hypothesis testing in a signal-plus-noise model via GLRT

To set a maximum false alarm rate (or Type I error) of $r > 0$ for large n, p, according to RMT, one must choose a threshold $f(\alpha)$ for T_p:

$$\mathbb{P}(T_p \geq f(\alpha)) = r \Leftrightarrow \mu_{\mathrm{TW}_1}((-\infty, A_p]) = r, \quad A_p = (f(\alpha) - (1+\sqrt{c})^2)(1+\sqrt{c})^{-\frac{4}{3}} c^{\frac{1}{6}} n^{\frac{2}{3}} \tag{5}$$

with μ_{TW_1} the Tracy-Widom distribution in RMT.

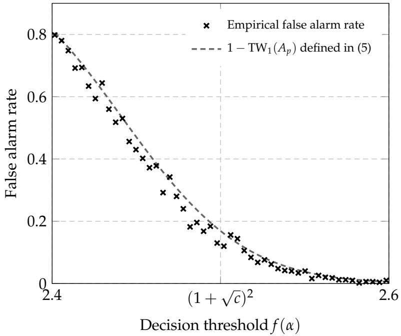


Figure: Comparison between empirical false alarm rates and $1 - \mathrm{TW}_1(A_p)$ for A_p of the form in (5), as a function of the threshold $f(\alpha) \in [(1+\sqrt{c})^2 - 5n^{-2/3}, (1+\sqrt{c})^2 + 5n^{-2/3}]$, for $p = 256$, $n = 1\,024$ and $\sigma = 1$.

“Curse of dimensionality”: loss of relevance of Euclidean distance

- Binary Gaussian mixture classification $\mathbf{x} \in \mathbb{R}^p$:

$$\mathcal{C}_1 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1), \text{ versus } \mathcal{C}_2 : \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_2, \mathbf{C}_2);$$

- Neyman-Pearson test: classification is possible only when

$$\boxed{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| \geq C_\mu, \text{ or } \|\mathbf{C}_1 - \mathbf{C}_2\| \geq C_{\mathbf{C}} \cdot p^{-1/2}}$$

for some constants $C_\mu, C_{\mathbf{C}} > 0$ [CLM18].

- In this non-trivial setting, for $\mathbf{x}_i \in \mathcal{C}_a, \mathbf{x}_j \in \mathcal{C}_b$:

$$\boxed{\max_{1 \leq i \neq j \leq n} \left\{ \frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_j\|^2 - \frac{2}{p} \operatorname{tr} \mathbf{C}^\circ \right\} \xrightarrow{a.s.} 0}$$

as $n, p \to \infty$ (i.e., $n \sim p$), for $\mathbf{C}^\circ \equiv \frac{1}{2}(\mathbf{C}_1 + \mathbf{C}_2)$, regardless of the classes $\mathcal{C}_a, \mathcal{C}_b$!

[0] Romain Couillet, Zhenyu Liao, and Xiaoyi Mai. “Classification asymptotics in the random matrix regime”. In: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE. 2018, pp. 1875–1879

Loss of relevance of Euclidean distance: visual representation

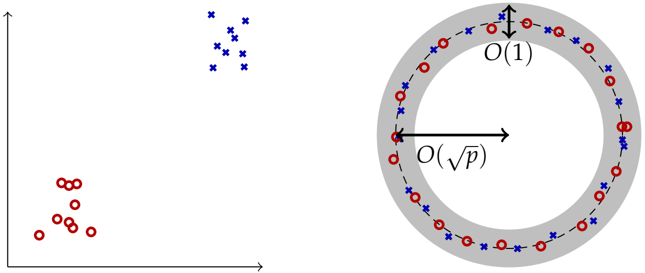
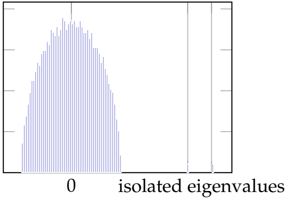


Figure: Visual representation of classification in (**left**) small and (**right**) large dimensions.

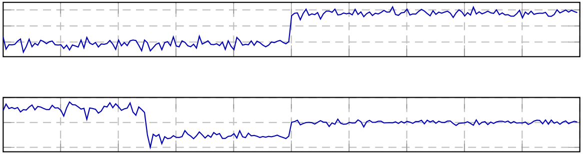
$\Rightarrow$ Direct consequence to various distance-based machine learning methods (e.g., kernel spectral clustering)!

Reminder on kernel spectral clustering

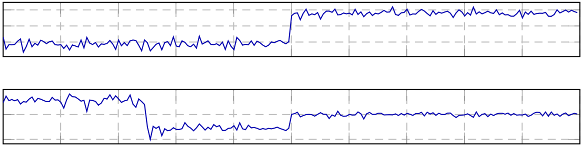
Two-step classification of n data points with distance kernel $\mathbf{K} \equiv \{f(\|\mathbf{x}_i - \mathbf{x}_j\|^2/p)\}_{i,j=1}^n$:



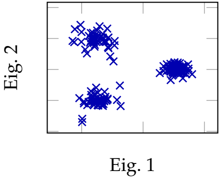
$\Downarrow$ **Top eigenvectors** $\Downarrow$



Reminder on kernel spectral clustering



$\Downarrow$ **K-dimensional representation** $\Downarrow$

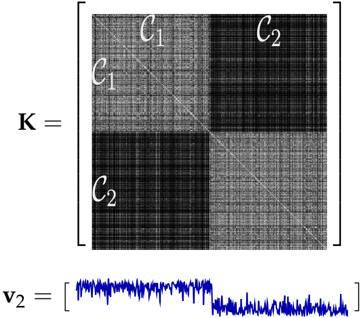


$\Downarrow$

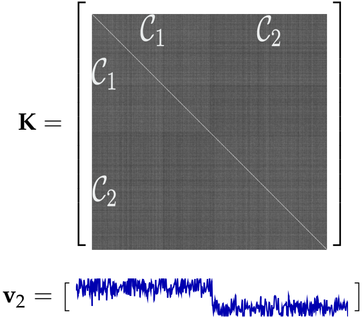
EM or k-means clustering

Cluster Gaussian data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbf{R}^p$ into $\mathcal{C}_1$ or $\mathcal{C}_2$, with second top eigenvectors $\mathbf{v}_2$ of heat kernel $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$, small and large dimensional data.

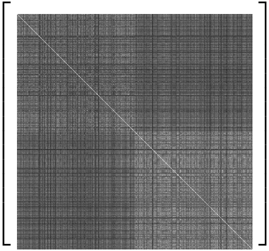
(a) $p = 5, n = 500$

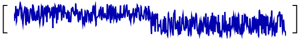


(b) $p = 250, n = 500$

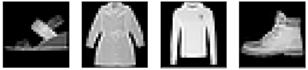


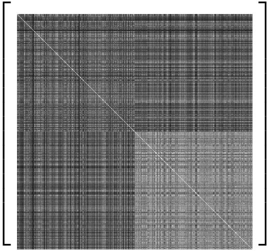
(a) MNIST

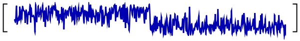
$\mathbf{K} =$ 

$\mathbf{v}_2 =$ 

(b) Fashion-MNIST



$\mathbf{K} =$ 

$\mathbf{v}_2 =$ 

- "local" linearization of *nonlinear* kernel matrices in large dimensions, e.g., Gaussian kernel matrix $\mathbf{K}_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2p)$ with $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{I}_p$ (e.g., $\mathcal{C}_1 : \mathbf{x}_i = \boldsymbol{\mu}_1 + \mathbf{z}_i$ versus $\mathcal{C}_2 : \mathbf{x}_j = \boldsymbol{\mu}_2 + \mathbf{z}_j$) so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2/p \xrightarrow{a.s.} 2, \text{ and } \mathbf{K} = \exp\left(-\frac{2}{2}\right)\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T} + \frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\right) + g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T} + * + o_{\|\cdot\|}(1)$$

  with Gaussian $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n] \in \mathbb{R}^{p\times n}$ and class-information $\mathbf{j} = [\mathbf{1}_{n/2}; -\mathbf{1}_{n/2}]$,
- accumulated effect of small "hidden" statistical information ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ in this case)

A RMT viewpoint of large kernel matrices

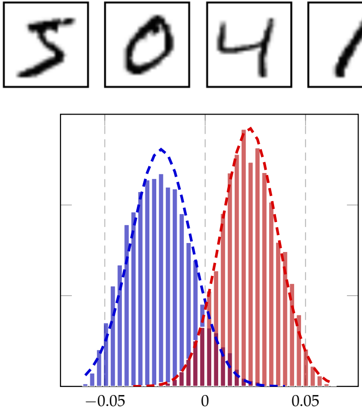
Therefore

- **entry-wise**:

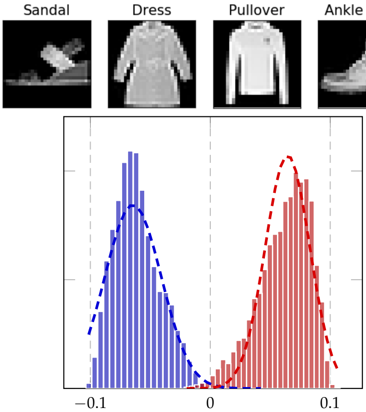
$$\mathbf{K}_{ij} = \exp(-1)\left(1 + \underbrace{\frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j}_{O(p^{-1/2})}\right) \pm \underbrace{\frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)}_{O(p^{-1})} + *, \text{ so that } \frac{1}{p}g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|) \ll \frac{1}{p}\mathbf{z}_i^\mathsf{T}\mathbf{z}_j,$$

- **spectrum-wise**:
  - $\|\mathbf{K} - \exp(-1)\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\| \not\to 0$;
  - $\|\frac{1}{p}\mathbf{Z}^\mathsf{T}\mathbf{Z}\| = O(1)$ and $\|g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\frac{1}{p}\mathbf{j}\mathbf{j}^\mathsf{T}\| = O(1)$!
- **Same** phenomenon as the sample covariance example: $[\hat{\mathbf{C}} - \mathbf{C}]_{ij} \to 0 \not\Rightarrow \|\hat{\mathbf{C}} - \mathbf{C}\| \to 0$!

$\Rightarrow$ With **RMT**, we understand kernel spectral clustering for large dimensional data!



(a) MNIST



(b) Fashion-MNIST

- (Strong) law of large numbers (LLN): for a sequence of i.i.d. random variables $x_1, \ldots, x_p$ with the same expectation $\mathbb{E}[x_i] = \mu$, we have

$$\frac{1}{p} \sum_{i=1}^{p} x_i \to \mu, \tag{6}$$

almost surely as $p \to \infty$.

- Central limit theorem (CLT, Lindeberg–Lévy tyep): for a sequence of i.i.d. random variables $x_1, \ldots, x_p$ with the same expectation $\mathbb{E}[x_i] = \mu$ and variance $\operatorname{Var}[x_i] = \sigma^2 < \infty$, we have

$$\sqrt{p} \left(\frac{1}{p} \sum_{i=1}^{p} (x_i - \mu) \right) \to \mathcal{N}(0, \sigma^2), \tag{7}$$

in distribution as $p \to \infty$.

OK with LLN and CLT, so what?

Different view of LLN and CLT: large-dimensional *deterministic* behavior and fluctuation.

Single scalar random variables

- *Scalar* random variable $x \in \mathbb{R}$, characterize its behavior distribution/law, characteristic function and/or successive moments, etc.
- x in general *not* expected to establish some kind of "close-to-deterministic" behavior.
- True for a *single observation*, although certainly the sum of many such random variables may concentrate and exhibit a close-to-deterministic behavior.

Random vectors: many scalar random variables

Consider a set of size p i.i.d. realizations/copies of such random variable. As a random vector $\mathbf{x} = [x_1, \ldots, x_p]^\mathsf{T} \in \mathbb{R}^p$, with $\mathbb{E}[x_i] = \mu$, $\mathrm{Var}[x_i] = 1$, $i \in \{1, \ldots, p\}$.

- as p independent *scalar* random variables $x \in \mathbb{R}$; or
- as a single realization of a *random vector* $\mathbf{x} \in \mathbb{R}^p$, having independent entries.

(i) **Scalar**: nothing more can be said about each *individual* random variable:

- inappropriate to predict the behavior of x_i with any *deterministic* value
- in general *incorrect* to say "the random x_i is close to $\mu = \mathbb{E}[x_i]$", since, for x_i with $\mathbb{E}[x] = \mu$ and $\mathrm{Var}[x] = 1$, by Chebyshev's inequality.

$$\mathbb{P}(|x - \mu| \geq t) \leq t^{-2}, \quad \forall t > 0. \tag{8}$$

- random fluctuation $x_i - \mathbb{E}[x_i]$ can be as large as $\mu = \mathbb{E}[x_i]$.

(ii) **Vector**: a different picture: single realization of random vector $\mathbf{x}/\sqrt{p} \in \mathbb{R}^p$.

- cannot say anything in general about each individual vector $\mathbf{x}$.
- however, if we are interested in only the (scalar and linear) observations of the random vector $\mathbf{x}/\sqrt{p} \in \mathbb{R}^p$ (with $\mathbb{E}[\mathbf{x}] = \mu \mathbf{1}_p/\sqrt{p}$), we known much more:

$$\frac{1}{p}\mathbf{x}^\mathsf{T}\mathbf{1}_p \xrightarrow{a.s.} \mathbb{E}[x_i] = \mu, \quad \frac{1}{\sqrt{p}}(\mathbf{x} - \mu\mathbf{1}_p)^\mathsf{T}\mathbf{1}_p \xrightarrow{d} \mathcal{N}(0,1), \quad p \to \infty. \tag{9}$$

OK with LLN and CLT, so what?

This is

$$\frac{1}{p}\mathbf{x}^\mathsf{T}\mathbf{1}_p \simeq \underbrace{\mu}_{O(1)} + \underbrace{\frac{1}{\sqrt{p}}\mathcal{N}(0,1)}_{O(p^{-1/2})}.$$

- a large dimensional random vector $\mathbf{x}/\sqrt{p} \in \mathbb{R}^p$, when "observed" via the linear map $\mathbf{1}_p^\mathsf{T}(\cdot)/\sqrt{p}$ of unit Euclidean norm (i.e., of "scale" independent of p);
- leads to $\mathbf{x}$ (when "observed" in this way) exhibiting the joint behavior of:
  - (i) approximately, in its first order, a *deterministic* quantity μ; and
  - (ii) in its second-order, a **universal** Gaussian fluctuation that is **strongly concentrated** and **independent** of the specific law of x_i.

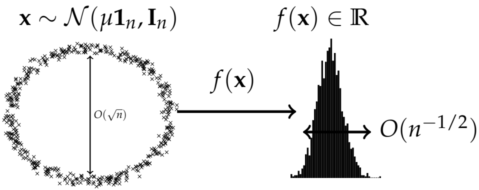


Figure: (**Left**) A "visualization" of independent realizations of $\mathbf{x} \sim \mathcal{N}(\mu\mathbf{1}_n, \mathbf{I}_n)$ with $n = 100$. (**Right**) Concentration behavior of scalar observations $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{1}_n/n$.

What about random matrices?

- As in the case of (high-dimensional) random vectors, we should **NOT** expect random matrices themselves converge in any useful sense;
- e.g., there does **NOT** exist deterministic matrix $\bar{\mathbf{X}}$ so that the random matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$

$$\|\mathbf{X} - \bar{\mathbf{X}}\| \to 0, \tag{10}$$

  in spectral norm as $p \to \infty$ (in probability or almost surely);
- nonetheless, "properly scaled" scalar observations $f: \mathbb{R}^{p \times p} \to \mathbb{R}$ of $\mathbf{X}$ **DO** converge, and there exists deterministic $\bar{\mathbf{X}}$ such that

$$f(\mathbf{X}) - f(\bar{\mathbf{X}}) \to 0, \tag{11}$$

  as $p \to \infty$. We say such $\bar{\mathbf{X}}$ is a **deterministic equivalent** of the random matrix $\mathbf{X}$.
- observation f of interest in RMT include (empirical) eigenvalue measure, linear spectral statistics (LSS), specific eigenvalue location, projection of eigenvectors, etc.

Deterministic equivalent for RMT: intuition and a few words on the proof

What is actually happening with scalar observations of random matrices and the deterministic equivalent (DE)?

- while the random matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$ remains random as the dimension p grows (in fact even "more" random due to the growing degrees of freedom);
- scalar observation $f(\mathbf{X})$ of $\mathbf{X}$ becomes "more concentrated" as $p \to \infty$;
  - the random $f(\mathbf{X})$, if concentrates, must concentrated around its expectation $\mathbb{E}[f(\mathbf{X})]$;
  - in fact, as $p \to \infty$, more randomness in $\mathbf{X} \Rightarrow \operatorname{Var}[f(\mathbf{X})] \to 0$, e.g., $\operatorname{Var}[f(\mathbf{X})] = p^{-4}$;
  - if the functional $f \colon \mathbb{R}^{p \times p} \to \mathbb{R}$ is linear, then $\mathbb{E}[f(\mathbf{X})] = f(\mathbb{E}[\mathbf{X}])$.
- So, to propose a DE, it suffices to evaluate $\mathbb{E}[\mathbf{X}]$:
  - **however**, $\mathbb{E}[\mathbf{X}]$ may be hardly accessible (due to integration)
  - find a simple and more accessible deterministic $\bar{\mathbf{X}}$ with $\bar{\mathbf{X}} \simeq \mathbb{E}[\mathbf{X}]$ in some sense for p large, e.g., $\|\bar{\mathbf{X}} - \mathbb{E}[\mathbf{X}]\| \to 0$ as $p \to \infty$; and
  - show variance of $f(\mathbf{X})$ decay sufficiently fast as $p \to \infty$.
- We say $\bar{\mathbf{X}}$ is a DE for $\mathbf{X}$ when $f(\mathbf{X})$ is evaluated, and denote $\mathbf{X} \leftrightarrow \bar{\mathbf{X}}$.

Fundamental Objects

Core interest of RMT: evaluation of eigenvalues and eigenvectors of a random matrix.

Definition (Resolvent)

For a symmetric/Hermitian matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$, the resolvent $\mathbf{Q}_\mathbf{X}(z)$ of $\mathbf{X}$ is defined, for $z \in \mathbb{C}$ not an eigenvalue of $\mathbf{X}$, as $\mathbf{Q}_\mathbf{X}(z) \equiv \left(\mathbf{X} - z\mathbf{I}_p\right)^{-1}$.

Definition (Empirical spectral measure)

For symmetric $\mathbf{X} \in \mathbb{R}^{p \times p}$, the *empirical spectral measure/distribution (ESD)* $\mu_\mathbf{X}$ of $\mathbf{X}$ is defined as the normalized counting measure of the eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_p(\mathbf{X})$ of $\mathbf{X}$, i.e., $\mu_\mathbf{X} \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(\mathbf{X})}$, where δ_x represents the Dirac measure at x.

Resolvent as the core object

Objects of interest	Functionals of resolvent $\mathbf{Q}_{\mathbf{X}}(z)$
Empirical spectral measure $\mu_{\mathbf{X}}$ of $\mathbf{X}$	Stieltjes transform $m_{\mu_{\mathbf{X}}}(z) = \frac{1}{p} \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z)$
Linear spectral statistics (LSS): $f(\mathbf{X}) \equiv \frac{1}{p} \sum_i f(\lambda_i(\mathbf{X}))$	Integration of trace of $\mathbf{Q}_{\mathbf{X}}(z)$: $-\frac{1}{2\pi \imath} \oint_\Gamma f(z) \frac{1}{p} \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z)\, dz$ (via Cauchy's integral)
Projections of eigenvectors $\mathbf{v}^\mathsf{T} \mathbf{u}(\mathbf{X})$ and $\mathbf{v}^\mathsf{T} \mathbf{U}(\mathbf{X})$ onto some given vector $\mathbf{v} \in \mathbb{R}^p$	Bilinear form $\mathbf{v}^\mathsf{T} \mathbf{Q}_{\mathbf{X}}(z) \mathbf{v}$ of $\mathbf{Q}_{\mathbf{X}}$
General matrix functional $F(\mathbf{X}) = \sum_i f(\lambda_i(\mathbf{X})) \mathbf{v}_1^\mathsf{T} \mathbf{u}_i(\mathbf{X}) \mathbf{u}_i(\mathbf{X})^\mathsf{T} \mathbf{v}_2$ involving both eigenvalues and eigenvectors	Integration of bilinear form of $\mathbf{Q}_{\mathbf{X}}(z)$: $-\frac{1}{2\pi \imath} \oint_\Gamma f(z) \mathbf{v}_1^\mathsf{T} \mathbf{Q}_{\mathbf{X}}(z) \mathbf{v}_2\, dz$

Use resolvent for eigenvalue distribution

Definition (Resolvent)

For a symmetric/Hermitian matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$, the resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ of $\mathbf{X}$ is defined, for $z \in \mathbb{C}$ not an eigenvalue of $\mathbf{X}$, as $\mathbf{Q}_{\mathbf{X}}(z) \equiv \left(\mathbf{X} - z\mathbf{I}_p\right)^{-1}$.

Let $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$ be the spectral decomposition of $\mathbf{X}$, with $\mathbf{\Lambda} = \{\lambda_i(\mathbf{X})\}_{i=1}^p$ eigenvalues and $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{p \times p}$ the associated eigenvectors. Then,

$$\mathbf{Q}(z) = \mathbf{U}(\mathbf{\Lambda} - z\mathbf{I}_p)^{-1}\mathbf{U}^\mathsf{T} = \sum_{i=1}^p \frac{\mathbf{u}_i \mathbf{u}_i^\mathsf{T}}{\lambda_i(\mathbf{X}) - z}. \tag{12}$$

Thus, for $\mu_{\mathbf{X}} \equiv \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(\mathbf{X})}$ the ESD of $\mathbf{X}$,

$$\frac{1}{p}\operatorname{tr}\mathbf{Q}(z) = \frac{1}{p}\sum_{i=1}^p \frac{1}{\lambda_i(\mathbf{X}) - z} = \int \frac{\mu_{\mathbf{X}}(dt)}{t - z}. \tag{13}$$

The Stieltjes transform

Definition (Stieltjes transform)

For a real probability measure μ with support $\operatorname{supp}(\mu)$, the *Stieltjes transform* $m_\mu(z)$ is defined, for all $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$, as

$$m_\mu(z) \equiv \int \frac{\mu(dt)}{t - z}. \tag{14}$$

For m_μ the Stieltjes transform of a probability measure μ, then

- m_μ is complex analytic on its domain of definition $\mathbb{C} \setminus \operatorname{supp}(\mu)$;
- it is bounded $|m_\mu(z)| \leq 1/\operatorname{dist}(z, \operatorname{supp}(\mu))$;
- it satisfies $m_\mu(z) > 0$ for $z < \inf \operatorname{supp}(\mu)$, $m_\mu(z) < 0$ for $z > \sup \operatorname{supp}(\mu)$ and $\Im[z] \cdot \Im[m_\mu(z)] > 0$ if $z \in \mathbb{C} \setminus \mathbb{R}$; and
- it is an increasing function on all connected components of its restriction to $\mathbb{R} \setminus \operatorname{supp}(\mu)$ (since $m'_\mu(x) = \int (t-x)^{-2} \mu(dt) > 0$) with $\lim_{x \to \pm\infty} m_\mu(x) = 0$ if $\operatorname{supp}(\mu)$ is bounded.

The inverse Stieltjes transform

Definition (Inverse Stieltjes transform)

For a, b continuity points of the probability measure μ, we have

$$\mu([a,b]) = \frac{1}{\pi} \lim_{y \downarrow 0} \int_a^b \Im \left[m_\mu(x + \imath y) \right] \, dx. \tag{15}$$

Besides, if μ admits a density f at x (i.e., $\mu(x)$ is differentiable in a neighborhood of x and $\lim_{\epsilon \to 0} (2\epsilon)^{-1} \mu([x - \epsilon, x + \epsilon]) = f(x)$),

$$f(x) = \frac{1}{\pi} \lim_{y \downarrow 0} \Im \left[m_\mu(x + \imath y) \right]. \tag{16}$$

Workflow: random matrix $\mathbf{X}$ of interest $\Rightarrow$ resolvent $\mathbf{Q}_\mathbf{X}(z)$ and ST $\frac{1}{p} \operatorname{tr} \mathbf{Q}_\mathbf{X}(z) = m_\mathbf{X}(z)$ $\Rightarrow$ study the limiting ST $m_\mathbf{X}(z) \to m(z)$ $\Rightarrow$ inverse ST to get limiting $\mu_\mathbf{X} \to \mu$.

Use the resolvent for eigenvalue functionals

Definition (Linear Spectral Statistic, LSS)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$, the *linear spectral statistics* (LSS) $f_{\mathbf{X}}$ of $\mathbf{X}$ is defined as the averaged statistics of the eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_p(\mathbf{X})$ of $\mathbf{X}$ via some function $f \colon \mathbb{R} \to \mathbb{R}$, that is

$$f(\mathbf{X}) = \frac{1}{p} \sum_{i=1}^{p} f(\lambda_i(\mathbf{X})) = \int f(t) \mu_{\mathbf{X}}(dt), \tag{17}$$

for $\mu_{\mathbf{X}}$ the ESD of $\mathbf{X}$.

Cauchy's integral formula

Theorem (Cauchy's integral formula)

For $\Gamma \subset \mathbb{C}$ a positively (i.e., counterclockwise) oriented simple closed curve and a complex function $f(z)$ analytic in a region containing Γ and its inside, then

(i) *if $z_0 \in \mathbb{C}$ is enclosed by Γ, $f(z_0) = -\frac{1}{2\pi\imath} \oint_\Gamma \frac{f(z)}{z_0 - z}\, dz$;*

(ii) *if not, $\frac{1}{2\pi\imath} \oint_\Gamma \frac{f(z)}{z_0 - z}\, dz = 0$.*

LSS via contour integration: For $\lambda_1(\mathbf{X}), \ldots, \lambda_p(\mathbf{X})$ eigenvalues of a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p \times p}$, some function $f \colon \mathbb{R} \to \mathbb{R}$ that is complex analytic in a compact neighborhood of the support $\operatorname{supp}(\mu_{\mathbf{X}})$ (of the ESD $\mu_{\mathbf{X}}$ of $\mathbf{X}$), then

$$f(\mathbf{X}) = \int f(t)\mu_{\mathbf{X}}(dt) = -\int \frac{1}{2\pi\imath} \oint_\Gamma \frac{f(z)\, dz}{t - z} \mu_{\mathbf{X}}(dt) = -\frac{1}{2\pi\imath} \oint_\Gamma f(z) m_{\mu_{\mathbf{X}}}(z)\, dz, \tag{18}$$

for *any* contour Γ that encloses $\operatorname{supp}(\mu_{\mathbf{X}})$, i.e., all the eigenvalues $\lambda_i(\mathbf{X})$.

LSS to retrieve the inverse Stieltjes transform formula

$$\frac{1}{p}\sum_{\lambda_i(\mathbf{X})\in[a,b]}\delta_{\lambda_i(\mathbf{X})} = -\frac{1}{2\pi\imath}\oint_\Gamma 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz$$

$$= -\frac{1}{2\pi\imath}\int_{a-\varepsilon_x-\imath\varepsilon_y}^{b+\varepsilon_x-\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz - \frac{1}{2\pi\imath}\int_{b+\varepsilon_x+\imath\varepsilon_y}^{a-\varepsilon_x+\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz$$

$$- \frac{1}{2\pi\imath}\int_{a-\varepsilon_x+\imath\varepsilon_y}^{a-\varepsilon_x-\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz - \frac{1}{2\pi\imath}\int_{b+\varepsilon_x-\imath\varepsilon_y}^{b+\varepsilon_x+\imath\varepsilon_y} 1_{\Re[z]\in[a-\varepsilon,b+\varepsilon]}(z)m_{\mu_\mathbf{X}}(z)\,dz.$$

- Since $\Re[m(x+\imath y)] = \Re[m(x-\imath y)], \Im[m(x+\imath y)] = -\Im[m(x-\imath y)]$;
- we have $\int_{a-\varepsilon_x}^{b+\varepsilon_x} m_{\mu_\mathbf{X}}(x-\imath\varepsilon_y)\,dx + \int_{b+\varepsilon_x}^{a-\varepsilon_x} m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)dx = -2\imath\int_{a-\varepsilon_x}^{b+\varepsilon_x}\Im[m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)]\,dx$;
- and consequently $\mu([a,b]) = \frac{1}{p}\sum_{\lambda_i(\mathbf{X})\in[a,b]}\lambda_i(\mathbf{X}) = \frac{1}{\pi}\lim_{\varepsilon_y\downarrow 0}\int_a^b \Im[m_{\mu_\mathbf{X}}(x+\imath\varepsilon_y)]\,dx$.

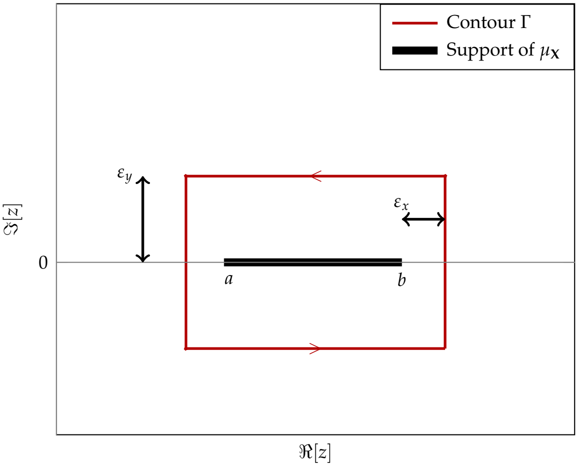


Figure: Illustration of a rectangular contour Γ and support of $\mu_{\mathbf{X}}$ on the complex plane.

Use resolvent for eigenvectors and eigenspace

Resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ contains eigenvector information about $\mathbf{X}$, recall

$$\mathbf{Q}_{\mathbf{X}}(z) = \sum_{i=1}^{p} \frac{\mathbf{u}_i \mathbf{u}_i^{\mathsf{T}}}{\lambda_i(\mathbf{X}) - z},$$

and that we have direct access to the i-th eigenvector $\mathbf{u}_i$ of $\mathbf{X}$ through

$$\mathbf{u}_i \mathbf{u}_i^{\mathsf{T}} = -\frac{1}{2\pi\imath} \oint_{\Gamma_{\lambda_i(\mathbf{X})}} \mathbf{Q}_{\mathbf{X}}(z)\, dz, \tag{19}$$

for $\Gamma_{\lambda_i(\mathbf{X})}$ a contour circling around $\lambda_i(\mathbf{X})$ only.

- seen as a matrix-version of LSS formula
- with the Stieltjes transform $m_{\mu_{\mathbf{X}}}(z)$ replaced by the associated resolvent $\mathbf{Q}_{\mathbf{X}}(z)$

Definition (Matrix spectral functionals)

For a symmetric matrix $\mathbf{X} \in \mathbb{R}^{p\times p}$, we say $F\colon \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ is a (matrix) spectral functional of $\mathbf{X}$,

$$F(\mathbf{X}) = \sum_{i\in\mathcal{I}\subseteq\{1,\ldots,p\}} f(\lambda_i(\mathbf{X}))\mathbf{u}_i\mathbf{u}_i^\mathsf{T}, \quad \mathbf{X} = \sum_{i=1}^{p} \lambda_i(\mathbf{X})\mathbf{u}_i\mathbf{u}_i^\mathsf{T}. \tag{20}$$

Spectral functional via contour integration: For $\mathbf{X} \in \mathbb{R}^{p\times p}$, resolvent $\mathbf{Q}_\mathbf{X}(z) = (\mathbf{X} - z\mathbf{I}_p)^{-1}$, $z \in \mathbb{C}$, and $f\colon \mathbb{R} \to \mathbb{R}$ analytic in a neighborhood of the contour $\Gamma_\mathcal{I}$ that circles around the eigenvalues $\lambda_i(\mathbf{X})$ of $\mathbf{X}$ with their indices in the set $\mathcal{I} \subseteq \{1,\ldots,p\}$,

$$F(\mathbf{X}) = -\frac{1}{2\pi\imath}\oint_{\Gamma_\mathcal{I}} f(z)\mathbf{Q}_\mathbf{X}(z)\,dz. \tag{21}$$

Example: eigenvector projection $(\mathbf{v}^\mathsf{T}\mathbf{u}_i)^2 = -\frac{1}{2\pi\imath}\oint_{\Gamma_{\lambda_i(\mathbf{X})}} \mathbf{v}^\mathsf{T}\mathbf{Q}_\mathbf{X}(z)\mathbf{v}\,dz$.

Sample covariance matrix in the large n, p regime

- **Problem**: estimate covariance $\mathbf{C} \in \mathbb{R}^{p \times p}$ from n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$,
- Maximum likelihood sample covariance matrix with entry-wise convergence

$$\hat{\mathbf{C}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \in \mathbb{R}^{p \times p}, \quad [\hat{\mathbf{C}}]_{ij} \to [\mathbf{C}]_{ij}$$

almost surely as $n \to \infty$: optimal for $n \gg p$ (or, for p "small").

- In the regime $n \sim p$, conventional wisdom breaks down: for $\mathbf{C} = \mathbf{I}_p$ with $n < p$, $\hat{\mathbf{C}}$ has at least $p - n$ zero eigenvalues:

$$\boxed{\|\hat{\mathbf{C}} - \mathbf{C}\| \not\to 0, \quad n, p \to \infty} \Rightarrow \text{ eigenvalue mismatch and not consistent!}$$

- due to $\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\| \leq p\|\mathbf{A}\|_\infty$ for $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\|\mathbf{A}\|_\infty \equiv \max_{ij} |\mathbf{A}_{ij}|$.

When is one in the random matrix regime? Almost always!

What about $n = 100p$? For $\mathbf{C} = \mathbf{I}_p$, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$: MP law

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

where $E_- = (1 - \sqrt{c})^2$, $E_+ = (1 + \sqrt{c})^2$ and $(x)^+ \equiv \max(x, 0)$. Close match!

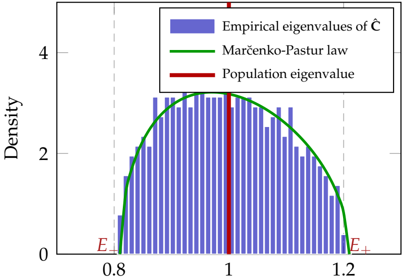


Figure: Eigenvalue distribution of $\hat{\mathbf{C}}$ versus Marčenko-Pastur law, $p = 500$, $n = 50\,000$.

- eigenvalues span on $[E_- = (1 - \sqrt{\mathbf{c}})^2, E_+ = (1 + \sqrt{\mathbf{c}})^2]$.
- for $\mathbf{n = 100p}$, on a range of $\pm 2\sqrt{\mathbf{c}} = \pm \mathbf{0.2}$ around the population eigenvalue $\mathbf{1}$.

Theorem (Marčenko–Pastur law)

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix with i.i.d. entries of zero mean and unit variance. Denote $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ the resolvent of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{22}$$

with $m(z)$ the unique Stieltjes transform solution to

$$zcm^2(z) - (1 - c - z)m(z) + 1 = 0. \tag{23}$$

Moreover, the empirical spectral measure $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ converges weakly to the probability measure μ

$$\boxed{\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi c x}\sqrt{(x - E_-)^+ (E_+ - x)^+}\, dx,} \tag{24}$$

where $E_\pm = (1 \pm \sqrt{c})^2$ and $(x)^+ = \max(0, x)$, known as the Marčenko-Pastur law.

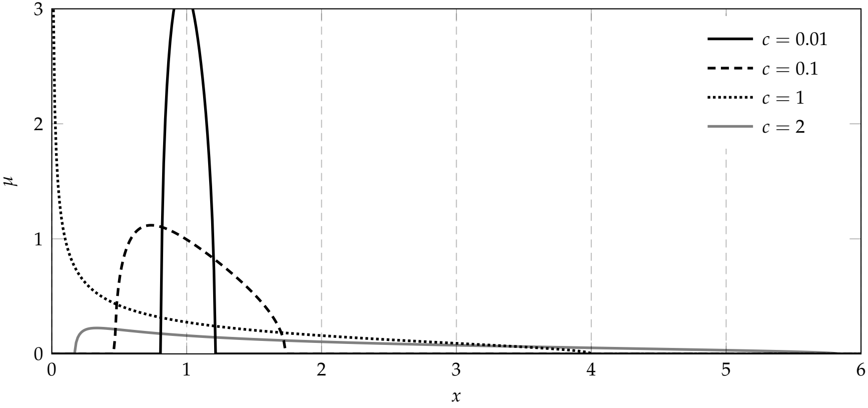


Figure: Marčenko-Pastur distribution for different values of c.

Proof of Marčenko–Pastur law

Workflow: random matrix $\mathbf{X}$ of interest $\Rightarrow$ resolvent $\mathbf{Q}_{\mathbf{X}}(z)$ and ST $\frac{1}{p} \operatorname{tr} \mathbf{Q}_{\mathbf{X}}(z) = m_{\mathbf{X}}(z)$
$\Rightarrow$ study the limiting ST $m_{\mathbf{X}}(z) \to m(z) \Rightarrow$ inverse ST to get limiting $\mu_{\mathbf{X}} \to \mu$.

Definition (Empirical Spectral Distribution, ESD)

For symmetric $\mathbf{X} \in \mathbb{R}^{p \times p}$, the *empirical spectral distribution (ESD)* $\mu_{\mathbf{X}}$ of $\mathbf{X}$ is defined as the normalized counting measure of the eigenvalues $\lambda_1(\mathbf{X}), \ldots, \lambda_p(\mathbf{X})$ of $\mathbf{X}$, i.e., $\mu_{\mathbf{X}} \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(\mathbf{X})}$, where δ_x represents the Dirac measure at x.

Definition (Stieltjes transform)

For a real probability measure μ with support $\operatorname{supp}(\mu)$, the *Stieltjes transform* $m_\mu(z)$ is defined, for all $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$, as

$$m_\mu(z) \equiv \int \frac{\mu(dt)}{t - z}. \tag{25}$$

Heuristic proof of MP law via "leave-one-out" approach

- "guess" $\bar{\mathbf{Q}}(z) = \mathbf{F}^{-1}(z)$ for some $\mathbf{F}(z)$ such that $\mathbb{E}[\mathbf{Q}] \simeq \bar{\mathbf{Q}}$ and $\frac{1}{p} \operatorname{tr} \mathbf{Q}(z) \simeq \frac{1}{p} \operatorname{tr} \bar{\mathbf{Q}}(z)$.
- for $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$,

$$\begin{aligned}\mathbf{Q}(z) - \bar{\mathbf{Q}}(z) &= \mathbf{Q}(z) \left(\mathbf{F}(z) + z\mathbf{I}_p - \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} \right) \bar{\mathbf{Q}}(z) \\ &= \mathbf{Q}(z) \left(\mathbf{F}(z) + z\mathbf{I}_p - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \right) \bar{\mathbf{Q}}(z).\end{aligned}$$

- for $\bar{\mathbf{Q}}(z) \leftrightarrow \mathbf{Q}(z)$ a DE for $\mathbf{Q}(z)$, look for $\frac{1}{p} \operatorname{tr}(\mathbf{Q}(z) - \bar{\mathbf{Q}}(z)) \to 0$,

$$\frac{1}{p} \operatorname{tr}(\mathbf{F}(z) + z\mathbf{I}_p)\bar{\mathbf{Q}}(z)\mathbf{Q}(z) - \frac{1}{n} \sum_{i=1}^n \frac{1}{p} \mathbf{x}_i^\mathsf{T} \bar{\mathbf{Q}}(z)\mathbf{Q}(z)\mathbf{x}_i \to 0. \tag{26}$$

- $\mathbf{x}_i^\mathsf{T} \bar{\mathbf{Q}}(z)\mathbf{Q}(z)\mathbf{x}_i / p$ as a quadratic form close to a trace form independent of $\mathbf{x}_i$.
- cannot be applied directly as $\mathbf{Q}(z)$ depends on $\mathbf{x}_i$.

Objective: "guess" the form of $\bar{\mathbf{Q}}(z) = \mathbf{F}^{-1}(z)$ for some $\mathbf{F}(z)$ so that $\frac{1}{p} \operatorname{tr} \mathbf{Q}(z) \simeq \frac{1}{p} \operatorname{tr} \bar{\mathbf{Q}}(z)$.

- use Sherman–Morrison to write $\mathbf{Q}(z)\mathbf{x}_i = \frac{\mathbf{Q}_{-i}(z)\mathbf{x}_i}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}(z)\mathbf{x}_i}$,
- now $\mathbf{Q}_{-i}(z) = (\frac{1}{n}\sum_{j\neq i}\mathbf{x}_j\mathbf{x}_j^\mathsf{T} - z\mathbf{I}_p)^{-1}$ is independent of $\mathbf{x}_i$,
- quadratic form close to the trace:

$$\frac{1}{p}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}(z)\mathbf{Q}(z)\mathbf{x}_i = \frac{\frac{1}{p}\mathbf{x}_i^\mathsf{T}\bar{\mathbf{Q}}(z)\mathbf{Q}_{-i}(z)\mathbf{x}_i}{1+\frac{1}{n}\mathbf{x}_i^\mathsf{T}\mathbf{Q}_{-i}(z)\mathbf{x}_i} \simeq \frac{\frac{1}{p}\operatorname{tr}\bar{\mathbf{Q}}(z)\mathbf{Q}_{-i}(z)}{1+\frac{1}{n}\operatorname{tr}\mathbf{Q}_{-i}(z)}. \tag{27}$$

- So $\frac{1}{p}\operatorname{tr}(\mathbf{F}(z)+z\mathbf{I}_p)\bar{\mathbf{Q}}(z)\mathbf{Q}(z) \simeq \frac{\frac{1}{p}\operatorname{tr}\bar{\mathbf{Q}}(z)\mathbf{Q}(z)}{1+\frac{1}{n}\operatorname{tr}\mathbf{Q}(z)}$, and "guess" $\mathbf{F}(z) \simeq \left(-z+\frac{1}{1+\frac{1}{n}\operatorname{tr}\mathbf{Q}(z)}\right)\mathbf{I}_p$.
- self-consistent equation of limiting ST $m(z)$ as

$$\frac{1}{p}\operatorname{tr}\mathbf{Q}(z) \simeq m(z) = \frac{1}{-z+\frac{1}{1+\frac{p}{n}\frac{1}{p}\operatorname{tr}\mathbf{Q}(z)}} \simeq \frac{1}{-z+\frac{1}{1+\frac{p}{n}m(z)}}. \tag{28}$$

Heuristic proof of MP law via "leave-one-out"

Objective: "guess" the form of $\bar{\mathbf{Q}}(z) = \mathbf{F}^{-1}(z)$ for some $\mathbf{F}(z)$ $\frac{1}{p}\operatorname{tr}\mathbf{Q}(z) \simeq \frac{1}{p}\operatorname{tr}\bar{\mathbf{Q}}(z)$.

- we have $\mathbf{F}(z) = \left(-z + \frac{1}{1+\frac{1}{n}\operatorname{tr}\bar{\mathbf{Q}}(z)}\right)\mathbf{I}_p$,
- and $\bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p$ with $m(z)$ unique Stieltjes transform solution to

$$m(z) = \left(-z + \frac{1}{1+cm(z)}\right)^{-1}, \text{ or } zcm^2(z) - (1-c-z)m(z) + 1 = 0.$$

- has two solutions defined via the two values of the complex square root function (letting $z = \rho e^{\imath\theta}$ for $\rho \geq 0$ and $\theta \in [0, 2\pi)$, $\sqrt{z} \in \{\pm\sqrt{\rho}e^{\imath\theta/2}\}$)

$$m(z) = \frac{1-c-z}{2cz} + \frac{\sqrt{((1+\sqrt{c})^2 - z)((1-\sqrt{c})^2 - z)}}{2cz},$$

  only one of which is such that $\Im[z]\Im[m(z)] > 0$ by definition of Stieltjes transforms.
- apply inverse Stieltjes transform we conclude the proof.

Some thoughts on the “leave-one-out” proof

- **in essence**: propose $\bar{\mathbf{Q}}(z) \simeq \mathbb{E}[\mathbf{Q}(z)]$ (in spectral norm sense), but simple to evaluate (via a quadratic equation)
- quadratic form close to the trace: high-dimensional concentration (around the expectation), nothing more than LLN and concentration
- leave-one-out analysis of large-scale system: $\frac{1}{p} \operatorname{tr} \mathbf{Q}(z) \simeq \frac{1}{p} \operatorname{tr} \mathbf{Q}_{-i}(z)$ for n, p large.
- low complexity analysis of large random system: joint behavior of p eigenvalues $\overset{\text{RMT}}{\to}$ a single deterministic (quadratic) equation
- These are the main intuitions and ingredients for almost everything in RMT and high-dimensional statistics!
- **Side remark**: another more systematic and convenient RMT proof approach: “**Gaussian method**,” as the combination of Stein’s lemma (Gaussian integration by parts), Nash–Poincare inequality, and interpolation from Gaussian to non-Gaussian, see [CL22, Section 2.2.2] for details.

Wigner semicircle law

Theorem (Wigner semicircle law)

Let $\mathbf{X} \in \mathbb{R}^{n \times n}$ be symmetric and such that the $\mathbf{X}_{ij} \in \mathbb{R}$, $j \geq i$, are independent zero mean and unit variance random variables. Then, for $\mathbf{Q}(z) = (\mathbf{X}/\sqrt{n} - z\mathbf{I}_n)^{-1}$, as $n \to \infty$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_n, \tag{29}$$

with $m(z)$ the unique ST solution to

$$m^2(z) + zm(z) + 1 = 0. \tag{30}$$

The function $m(z)$ is the Stieltjes transform of the probability measure

$$\mu(dx) = \frac{1}{2\pi}\sqrt{(4 - x^2)^+}\, dx, \tag{31}$$

known as the Wigner semicircle law.

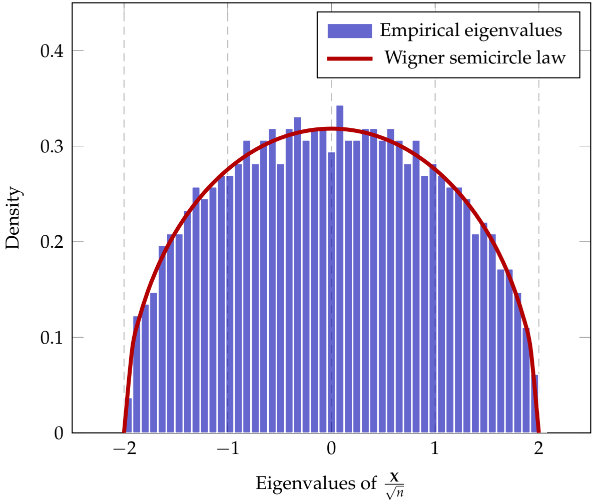


Figure: Histogram of the eigenvalues of $\mathbf{X}/\sqrt{n}$ versus Wigner semicircle law, for standard Gaussian $\mathbf{X}$ and $n = 1\,000$.

Theorem (General sample covariance matrix)

Let $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z} \in \mathbb{R}^{p\times n}$ *with nonnegative definite* $\mathbf{C} \in \mathbb{R}^{p\times p}$, $\mathbf{Z} \in \mathbb{R}^{p\times n}$ *having independent zero mean and unit variance entries. Then, as* $n, p \to \infty$ *with* $p/n \to c \in (0,\infty)$, *for* $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ *and* $\tilde{\mathbf{Q}}(z) = (\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X} - z\mathbf{I}_n)^{-1}$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = -\frac{1}{z}\left(\mathbf{I}_p + \tilde{m}_p(z)\mathbf{C}\right)^{-1}, \quad \tilde{\mathbf{Q}}(z) \leftrightarrow \bar{\tilde{\mathbf{Q}}}(z) = \tilde{m}_p(z)\mathbf{I}_n,$$

with $\tilde{m}_p(z)$ *unique solution to* $\tilde{m}_p(z) = \left(-z + \frac{1}{n}\operatorname{tr}\mathbf{C}\left(\mathbf{I}_p + \tilde{m}_p(z)\mathbf{C}\right)^{-1}\right)^{-1}$. *Moreover, if the empirical spectral measure of* $\mathbf{C}$ *converges* $\mu_{\mathbf{C}} \to \nu$ *as* $p \to \infty$, *then* $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}} \to \mu$, $\mu_{\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}} \to \tilde{\mu}$ *where* $\mu, \tilde{\mu}$ *admitting Stieltjes transforms* $m(z)$ *and* $\tilde{m}(z)$ *such that*

$$m(z) = \frac{1}{c}\tilde{m}(z) + \frac{1-c}{cz}, \quad \tilde{m}(z) = \left(-z + c\int \frac{t\nu(dt)}{1+\tilde{m}(z)t}\right)^{-1}. \tag{32}$$

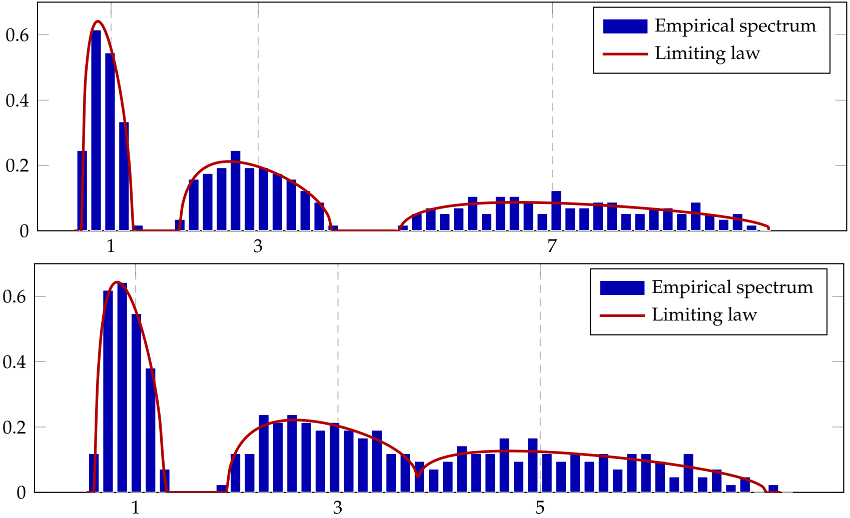
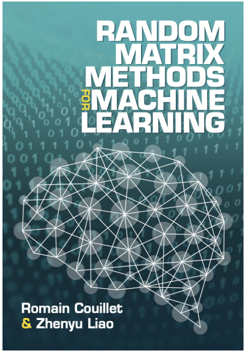


Figure: Histogram of the eigenvalues of $\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}$, $\mathbf{X} = \mathbf{C}^{\frac{1}{2}}\mathbf{Z} \in \mathbb{R}^{p\times n}$, $[\mathbf{Z}]_{ij} \sim \mathcal{N}(0,1)$, $n = 3\,000$; for $p = 300$ and $\mathbf{C}$ having spectral measure $\mu_{\mathbf{C}} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_7)$ **(top)** and $\mu_{\mathbf{C}} = \frac{1}{3}(\delta_1 + \delta_3 + \delta_5)$ **(bottle)**.

RMT for machine learning: from theory to practice!

Random matrix theory (RMT) for machine learning:

- **change of intuition** from small to large dimensional learning paradigm!
- **better understanding** of existing methods: why they work if they do, and what the issue is if they do not
- **improved novel methods** with performance guarantee!



- book "*Random Matrix Methods for Machine Learning*"
- by Romain Couillet and **Zhenyu Liao**
- Cambridge University Press, 2022
- a pre-production version of the book and exercise solutions at `https://zhenyu-liao.github.io/book/`
- MATLAB and Python codes to reproduce all figures at `https://github.com/Zhenyu-LIAO/RMT4ML`

Thank you! Q & A?