Tomorrow's large dimensional AI: renewed intuitions and new mathematics?

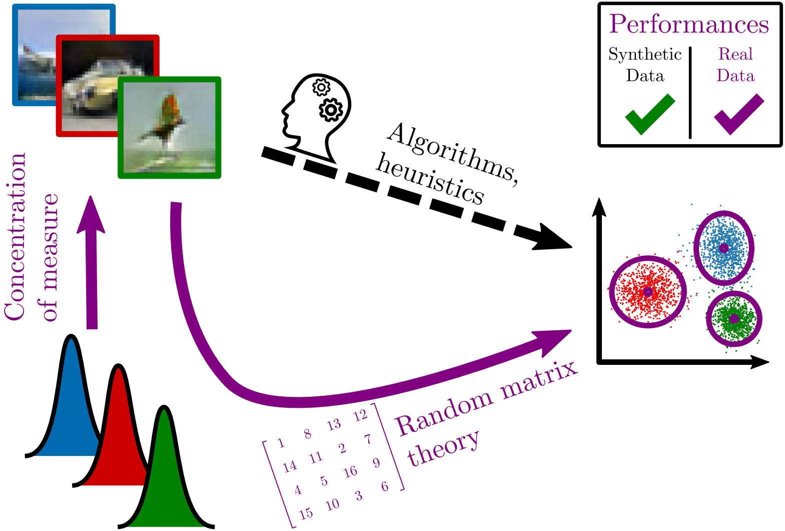
Workshop MACS COMET-SCA on "Automatics and AI"

Romain COUILLET

CentraleSupélec, L2S, University of ParisSaclay, France
GSTATS IDEX DataScience Chair, GIPSA-lab, University Grenoble–Alpes, France.

June 2, 2021

A long story short...



Concentration of measure

Algorithms, heuristics

Random matrix theory

$$\begin{bmatrix} 1 & 8 & 13 & 12 \\ 14 & 11 & 2 & 7 \\ 4 & 5 & 16 & 9 \\ 15 & 10 & 3 & 6 \end{bmatrix}$$

Performances

Synthetic Data	Real Data
✓	✓

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- Spiked Models

Application to Machine Learning

Outline

Basics of Random Matrix Theory

Motivation: Large Sample Covariance Matrices

Spiked Models

Application to Machine Learning

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- Spiked Models

Application to Machine Learning

Context

Baseline scenario: $y_1, \ldots, y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1] = 0$, $E[y_1 y_1^*] = C_p$:

Context

Baseline scenario: $y_1, \ldots, y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1] = 0$, $E[y_1 y_1^*] = C_p$:

- If $y_1 \sim \mathcal{N}(0, C_p)$, ML estimator for C_p is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^*$$

  ($Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}$).

Context

Baseline scenario: $y_1, \ldots, y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1] = 0$, $E[y_1 y_1^*] = C_p$:

- If $y_1 \sim \mathcal{N}(0, C_p)$, ML estimator for C_p is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^*$$

  ($Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}$).

- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_p \xrightarrow{\text{a.s.}} C_p.$$

  or equivalently, in spectral norm

$$\left\| \hat{C}_p - C_p \right\| \xrightarrow{\text{a.s.}} 0.$$

Context

Baseline scenario: $y_1, \ldots, y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1] = 0$, $E[y_1 y_1^*] = C_p$:

- If $y_1 \sim \mathcal{N}(0, C_p)$, ML estimator for C_p is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^*$$

  ($Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}$).

- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_p \xrightarrow{\text{a.s.}} C_p.$$

  or equivalently, in spectral norm

$$\left\| \hat{C}_p - C_p \right\| \xrightarrow{\text{a.s.}} 0.$$

Random Matrix Regime

- No longer valid if $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

$$\left\| \hat{C}_p - C_p \right\| \not\to 0.$$

Context

Baseline scenario: $y_1, \ldots, y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1] = 0$, $E[y_1 y_1^*] = C_p$:

- If $y_1 \sim \mathcal{N}(0, C_p)$, ML estimator for C_p is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^*$$

  ($Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}$).

- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_p \xrightarrow{\text{a.s.}} C_p.$$

  or equivalently, in spectral norm

$$\left\| \hat{C}_p - C_p \right\| \xrightarrow{\text{a.s.}} 0.$$

Random Matrix Regime

- No longer valid if $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

$$\left\| \hat{C}_p - C_p \right\| \not\to 0.$$

- For practical p, n with $p \simeq n$, leads to dramatically wrong conclusions

Context

Baseline scenario: $y_1, \ldots, y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1] = 0$, $E[y_1 y_1^*] = C_p$:

- If $y_1 \sim \mathcal{N}(0, C_p)$, ML estimator for C_p is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} Y_p Y_p^* = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^*$$

  ($Y_p = [y_1, \ldots, y_n] \in \mathbb{C}^{p \times n}$).

- If $n \to \infty$, then, strong law of large numbers

$$\hat{C}_p \xrightarrow{\text{a.s.}} C_p.$$

  or equivalently, in spectral norm

$$\left\| \hat{C}_p - C_p \right\| \xrightarrow{\text{a.s.}} 0.$$

Random Matrix Regime

- No longer valid if $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

$$\left\| \hat{C}_p - C_p \right\| \not\to 0.$$

- For practical p, n with $p \simeq n$, leads to dramatically wrong conclusions
- **Even for $p = n/100$.**

The Large Dimensional Fallacies

Setting: $y_i \in \mathbb{R}^p$ i.i.d., $y_1 \sim \mathcal{CN}(0, I_p)$

The Large Dimensional Fallacies

Setting: $y_i \in \mathbb{R}^p$ i.i.d., $y_1 \sim \mathcal{CN}(0, I_p)$

- assume $p = p(n)$ such that $p/n \to c > 1$

The Large Dimensional Fallacies

Setting: $y_i \in \mathbb{R}^p$ i.i.d., $y_1 \sim \mathcal{CN}(0, I_p)$

- assume $p = p(n)$ such that $p/n \to c > 1$
- then, joint point-wise convergence

$$\max_{1\leq i,j\leq p} \left| \left[\hat{C}_p - I_p\right]_{ij} \right| = \max_{1\leq i,j\leq p} \left| \frac{1}{n} \underbrace{Y_{j,\cdot} Y_{i,\cdot}^{\mathsf{T}}}_{\text{vectors of } \mathcal{N}(0,1) \text{ entries}} - \boldsymbol{\delta}_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

The Large Dimensional Fallacies

Setting: $y_i \in \mathbb{R}^p$ i.i.d., $y_1 \sim \mathcal{CN}(0, I_p)$

- assume $p = p(n)$ such that $p/n \to c > 1$
- then, joint point-wise convergence

$$\max_{1 \leq i,j \leq p} \left| \left[\hat{C}_p - I_p \right]_{ij} \right| = \max_{1 \leq i,j \leq p} \left| \frac{1}{n} \underbrace{Y_{j,\cdot} Y_{i,\cdot}^{\mathsf{T}}}_{\text{vectors of } \mathcal{N}(0,1) \text{ entries}} - \boldsymbol{\delta}_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

- however, eigenvalue mismatch

$$0 = \lambda_1(\hat{C}_p) = \ldots = \lambda_{p-n}(\hat{C}_p) \leq \lambda_{p-n+1}(\hat{C}_p) \leq \ldots \leq \lambda_p(\hat{C}_p)$$
$$1 = \lambda_1(I_p) = \ldots = \lambda_{p-n}(I_p) = \lambda_{p-n+1}(\hat{C}_p) = \ldots = \lambda_p(I_p)$$

The Large Dimensional Fallacies

Setting: $y_i \in \mathbb{R}^p$ i.i.d., $y_1 \sim \mathcal{CN}(0, I_p)$

- assume $p = p(n)$ such that $p/n \to c > 1$
- then, joint point-wise convergence

$$\max_{1 \leq i,j \leq p} \left| \left[\hat{C}_p - I_p \right]_{ij} \right| = \max_{1 \leq i,j \leq p} \left| \frac{1}{n} \underbrace{Y_{j,\cdot} Y_{i,\cdot}^{\mathsf{T}}}_{\text{vectors of } \mathcal{N}(0,1) \text{ entries}} - \boldsymbol{\delta}_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

- however, eigenvalue mismatch

$$0 = \lambda_1(\hat{C}_p) = \ldots = \lambda_{p-n}(\hat{C}_p) \leq \lambda_{p-n+1}(\hat{C}_p) \leq \ldots \leq \lambda_p(\hat{C}_p)$$
$$1 = \lambda_1(I_p) = \ldots = \lambda_{p-n}(I_p) = \lambda_{p-n+1}(\hat{C}_p) = \ldots = \lambda_p(I_p)$$

⇒ no convergence in spectral norm.

The Marčenko–Pastur law

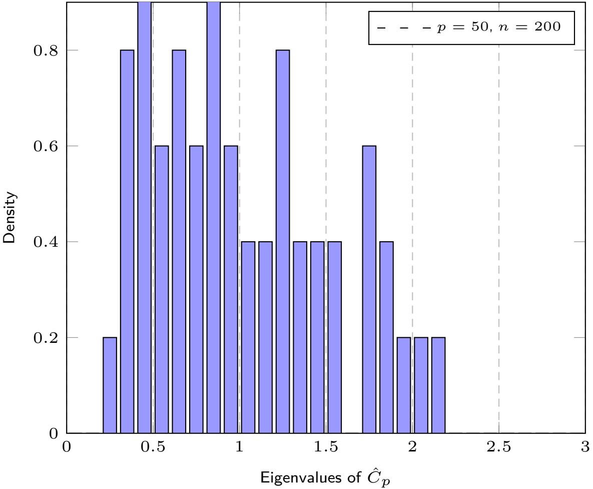


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

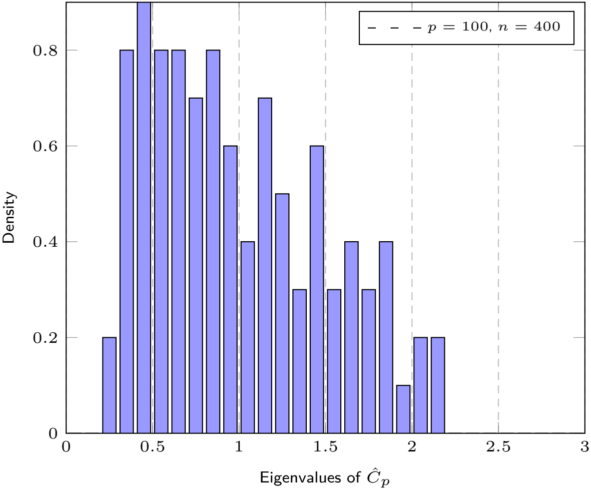


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

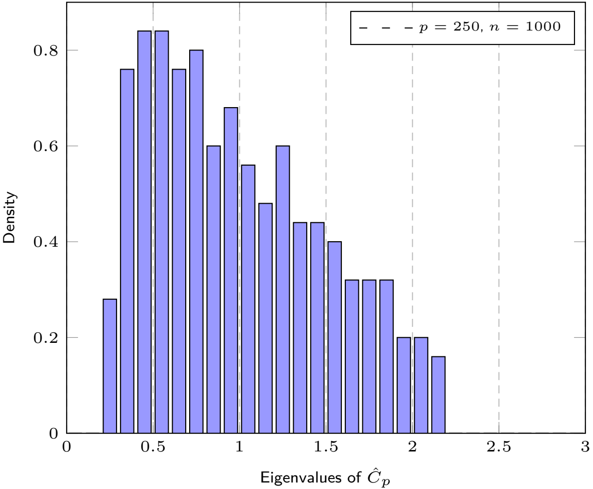


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

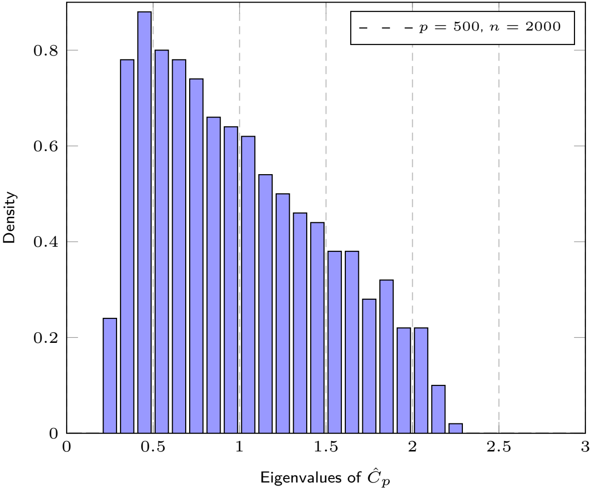


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

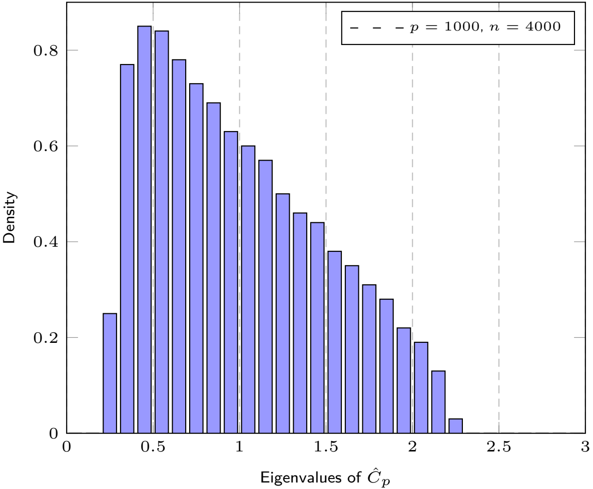


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

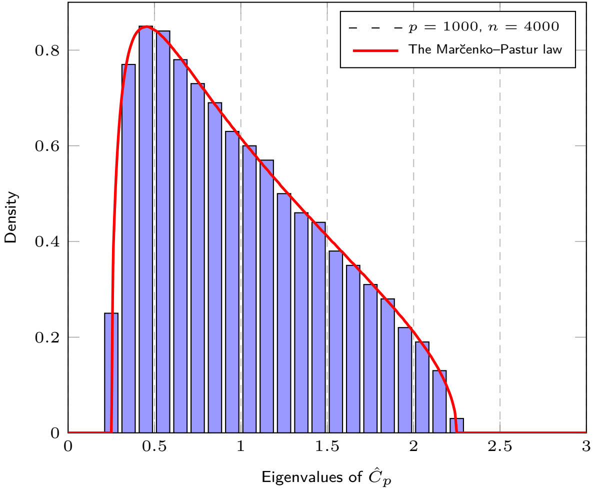


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $c = 1/4$, $C_p = I_p$.

The Marčenko–Pastur law

Definition (Empirical Spectral Density)

Empirical spectral density (e.s.d.) μ_p of Hermitian matrix $A_p \in \mathbb{C}^{p\times p}$ is

$$\mu_p = \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{\delta}_{\lambda_i(A_p)}.$$

The Marčenko–Pastur law

Definition (Empirical Spectral Density)

Empirical spectral density (e.s.d.) μ_p of Hermitian matrix $A_p \in \mathbb{C}^{p \times p}$ is

$$\mu_p = \frac{1}{p} \sum_{i=1}^{p} \boldsymbol{\delta}_{\lambda_i(A_p)}.$$

Theorem (Marčenko–Pastur Law **[Marčenko,Pastur'67]**)

$X_p \in \mathbb{C}^{p \times n}$ with i.i.d. zero mean, unit variance entries.
As $p, n \to \infty$ with $p/n \to c \in (0, \infty)$, e.s.d. μ_p of $\frac{1}{n} X_p X_p^*$ satisfies

$$\mu_p \xrightarrow{\text{a.s.}} \mu_c$$

weakly, where

- $\mu_c(\{0\}) = \max\{0, 1 - c^{-1}\}$

The Marčenko–Pastur law

Definition (Empirical Spectral Density)

Empirical spectral density (e.s.d.) μ_p of Hermitian matrix $A_p \in \mathbb{C}^{p\times p}$ is

$$\mu_p = \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{\delta}_{\lambda_i(A_p)}.$$

Theorem (Marčenko–Pastur Law **[Marčenko,Pastur'67]**)

$X_p \in \mathbb{C}^{p\times n}$ with i.i.d. zero mean, unit variance entries.
As $p, n \to \infty$ with $p/n \to c \in (0,\infty)$, e.s.d. μ_p of $\frac{1}{n}X_pX_p^*$ satisfies

$$\mu_p \xrightarrow{\text{a.s.}} \mu_c$$

weakly, where

- $\mu_c(\{0\}) = \max\{0, 1-c^{-1}\}$
- *on $(0,\infty)$, μ_c has continuous density f_c supported on $[(1-\sqrt{c})^2, (1+\sqrt{c})^2]$*

$$f_c(x) = \frac{1}{2\pi c x}\sqrt{(x-(1-\sqrt{c})^2)((1+\sqrt{c})^2-x)}.$$

The Marčenko–Pastur law

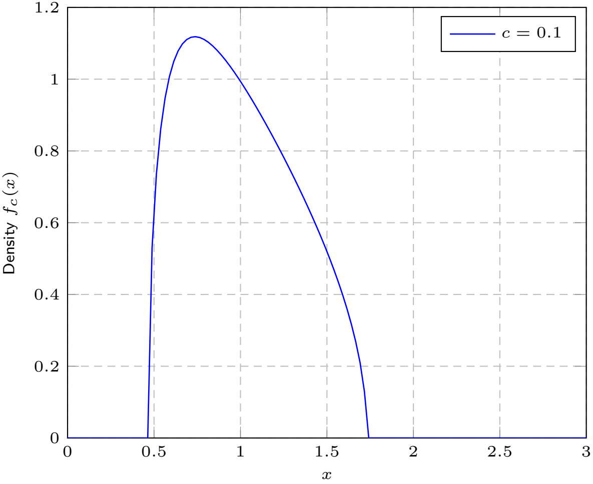


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

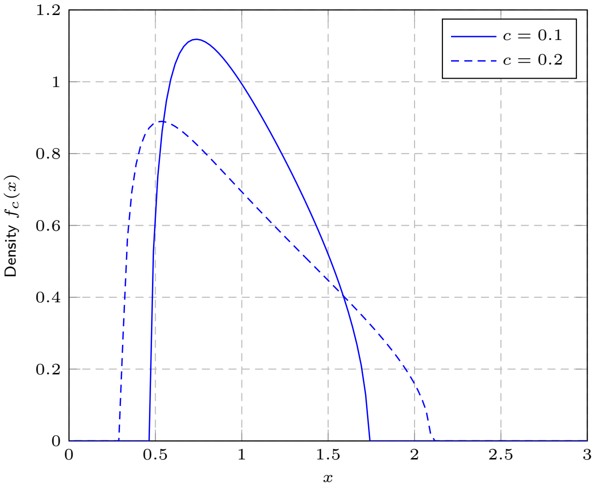


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

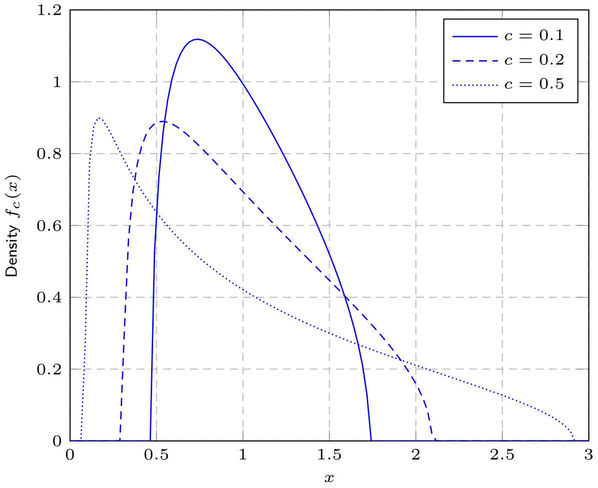


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

Outline

Basics of Random Matrix Theory

Motivation: Large Sample Covariance Matrices

Spiked Models

Application to Machine Learning

Spiked Models

Small rank perturbation: $C_p = I_p + P$, P of low rank.

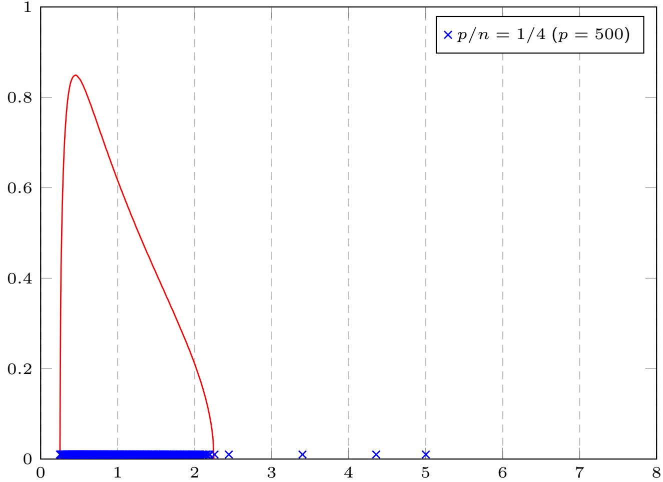


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^{\mathsf{T}}$, $\mathrm{eig}(C_p) = \{\underbrace{1, \ldots, 1}_{p-4}, 2, 3, 4, 5\}$.

Spiked Models

Small rank perturbation: $C_p = I_p + P$, P of low rank.

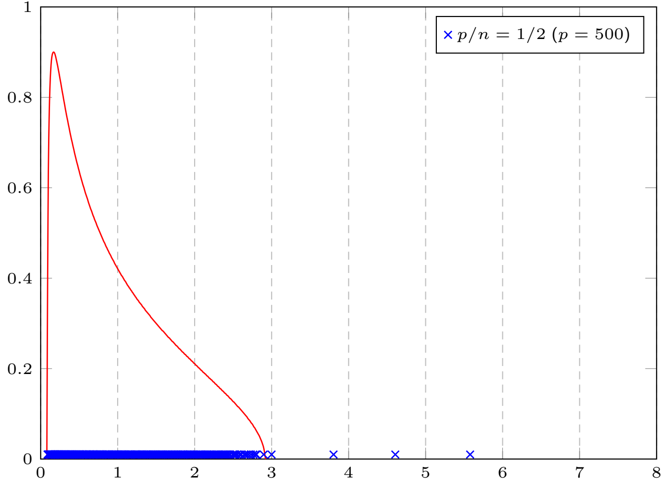


Figure: Eigenvalues of $\frac{1}{n}Y_pY_p^{\mathsf{T}}$, $\mathrm{eig}(C_p) = \{\underbrace{1,\ldots,1}_{p-4}, 2, 3, 4, 5\}$.

Spiked Models

Small rank perturbation: $C_p = I_p + P$, P of low rank.

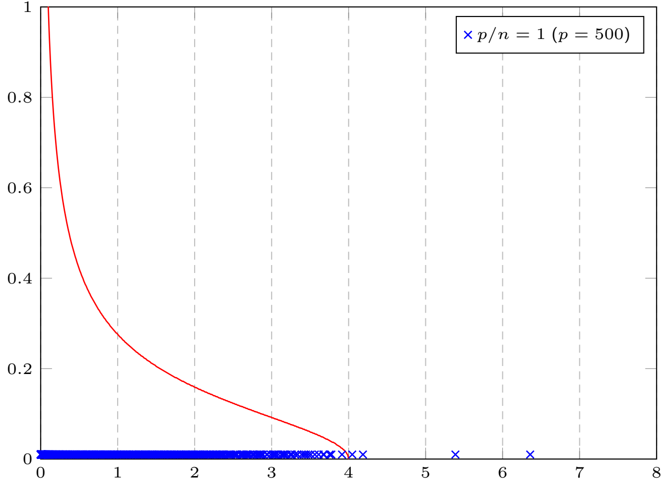


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^{\mathsf{T}}$, $\operatorname{eig}(C_p) = \{\underbrace{1, \ldots, 1}_{p-4}, 2, 3, 4, 5\}$.

Spiked Models

Small rank perturbation: $C_p = I_p + P$, P of low rank.

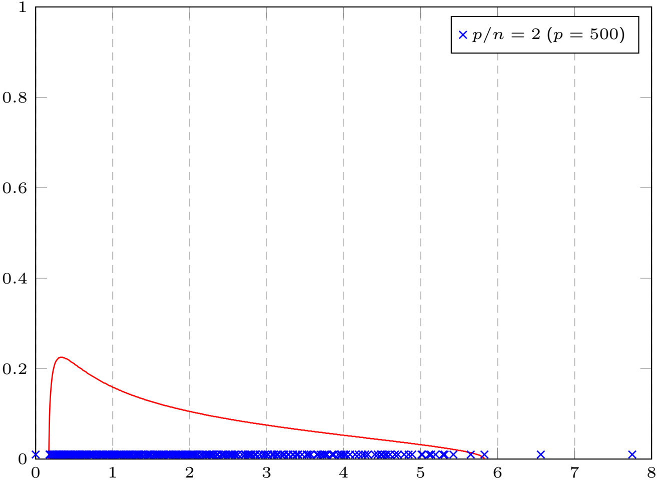


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^\mathsf{T}$, $\mathrm{eig}(C_p) = \{\underbrace{1, \ldots, 1}_{p-4}, 2, 3, 4, 5\}$.

Spiked Models

Theorem (Eigenvalues **[Baik,Silverstein'06]**)

Let $Y_p = C_p^{\frac{1}{2}} X_p$*, with*

- X_p *with i.i.d. zero mean, unit variance,* $E[|X_p|_{ij}^4] < \infty$*.*
- $C_p = I_p + P$*,* $P = U\Omega U^*$*, where, for* K *fixed,*

$$\Omega = \mathrm{diag}\left(\omega_1, \ldots, \omega_K\right) \in \mathbb{R}^{K\times K}, \text{ with } \omega_1 \geq \ldots \geq \omega_K > 0.$$

Spiked Models

Theorem (Eigenvalues **[Baik,Silverstein'06]**)

Let $Y_p = C_p^{\frac{1}{2}} X_p$, *with*

- X_p *with i.i.d. zero mean, unit variance,* $E[|X_p|_{ij}^4] < \infty$.
- $C_p = I_p + P$, $P = U\Omega U^*$, *where, for* K *fixed,*

$$\Omega = \mathrm{diag}\left(\omega_1, \ldots, \omega_K\right) \in \mathbb{R}^{K\times K}, \text{ with } \omega_1 \geq \ldots \geq \omega_K > 0.$$

Then, as $p, n \to \infty$, $p/n \to c \in (0, \infty)$, *denoting* $\lambda_m = \lambda_m(\frac{1}{n} Y_p Y_p^*)$ *(*$\lambda_m > \lambda_{m+1}$*),*

$$\lambda_m \xrightarrow{\text{a.s.}} \begin{cases} 1 + \omega_m + c\frac{1+\omega_m}{\omega_m} > (1+\sqrt{c})^2 & , \ \omega_m > \sqrt{c} \\ (1+\sqrt{c})^2 & , \ \omega_m \in (0, \sqrt{c}]. \end{cases}$$

Spiked Models

Theorem (Eigenvectors **[Paul'07]**)

Let $Y_p = C_p^{\frac{1}{2}} X_p$, *with*

- X_p *with i.i.d. zero mean, unit variance,* $E[|X_p|_{ij}^4] < \infty$.
- $C_p = I_p + P$, $P = U\Omega U^* = \sum_{i=1}^K \omega_i u_i u_i^*$, $\omega_1 > \ldots > \omega_M > 0$.

Spiked Models

Theorem (Eigenvectors **[Paul'07]**)

Let $Y_p = C_p^{\frac{1}{2}} X_p$, *with*

- X_p *with i.i.d. zero mean, unit variance,* $E[|X_p|_{ij}^4] < \infty$.
- $C_p = I_p + P$, $P = U\Omega U^* = \sum_{i=1}^K \omega_i u_i u_i^*$, $\omega_1 > \ldots > \omega_M > 0$.

Then, as $p, n \to \infty$, $p/n \to c \in (0, \infty)$, *for* $a, b \in \mathbb{C}^p$ *deterministic and* $\hat{u}_i$ *eigenvector of* $\lambda_i(\frac{1}{n} Y_p Y_p^*)$,

$$a^* \hat{u}_i \hat{u}_i^* b - \frac{1 - c\omega_i^{-2}}{1 + c\omega_i^{-1}} a^* u_i u_i^* b \cdot 1_{\omega_i > \sqrt{c}} \xrightarrow{\text{a.s.}} 0$$

In particular,

$$|\hat{u}_i^* u_i|^2 \xrightarrow{\text{a.s.}} \frac{1 - c\omega_i^{-2}}{1 + c\omega_i^{-1}} \cdot 1_{\omega_i > \sqrt{c}}.$$

Spiked Models

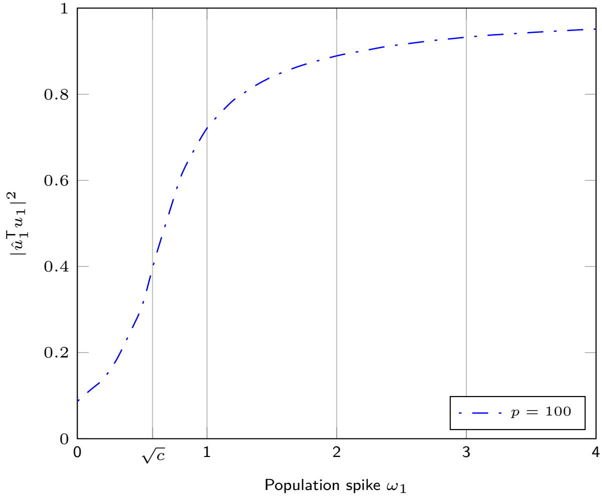


Figure: Simulated versus limiting $|\hat{u}_1^\mathsf{T} u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^\mathsf{T}$, $p/n = 1/3$, varying ω_1.

Spiked Models

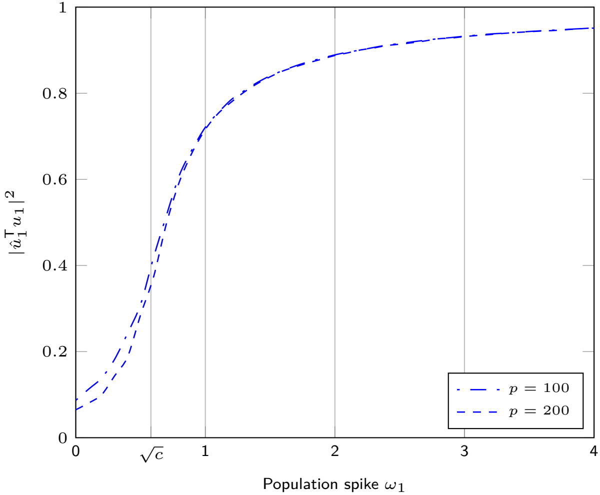


Figure: Simulated versus limiting $|\hat{u}_1^\mathsf{T} u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^\mathsf{T}$, $p/n = 1/3$, varying ω_1.

Spiked Models

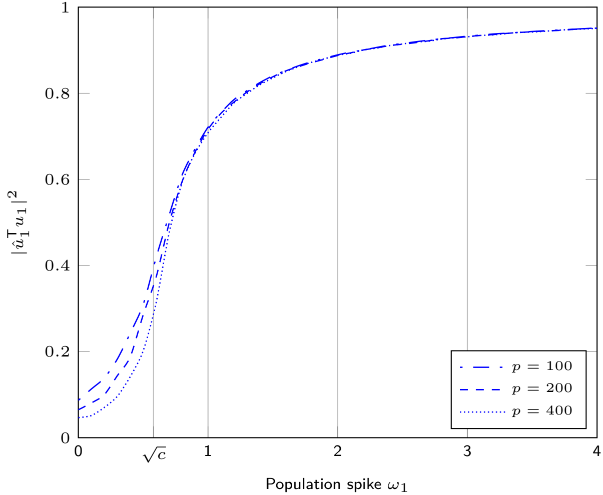


Figure: Simulated versus limiting $|\hat{u}_1^{\mathsf{T}} u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^{\mathsf{T}}$, $p/n = 1/3$, varying ω_1.

Spiked Models

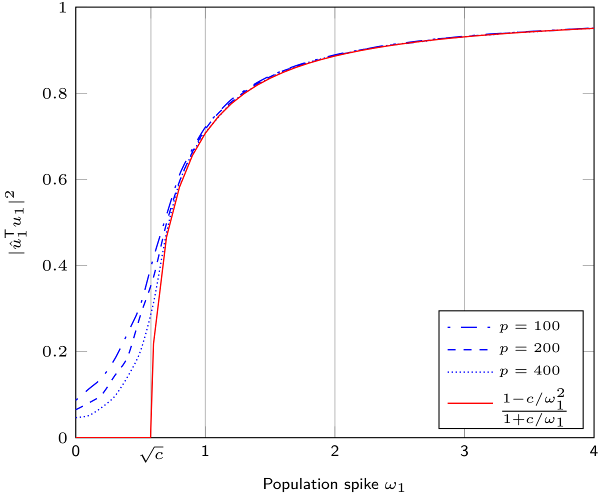


Figure: Simulated versus limiting $|\hat{u}_1^\mathsf{T} u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^\mathsf{T}$, $p/n = 1/3$, varying ω_1.

Other Spiked Models

Similar results for multiple matrix models with $\mathbb{E}[X_{ij}] = 0$ and P low rank:

- $K = \frac{1}{n}(I+P)^{\frac{1}{2}} X_p X_p^* (I+P)^{\frac{1}{2}}$
- $K = \frac{1}{n} X_p X_p^* + P$
- $K = \frac{1}{n} X_p^* (I+P) X$
- $K = \frac{1}{n}(X_p+P)^*(X_p+P)$
- etc.

Outline

Basics of Random Matrix Theory

- Motivation: Large Sample Covariance Matrices
- Spiked Models

Application to Machine Learning

Takeaway Message 1

“RMT Explains Why Machine Learning Intuitions
Collapse in Large Dimensions”