

# Random Matrix Advances in Large Dimensional Statistics, Machine Learning and Neural Nets

(EUSIPCO'2019, A Coruna, Spain)

Romain COUILLET, Malik TIOMOKO, Mohamed SEDDIK

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

September 2nd, 2019



CentraleSupélec



## Basics of Random Matrix Theory (**Romain COUILLET**)

- Motivation: Large Sample Covariance Matrices

- The Stieltjes Transform Method

- Spiked Models

- Other Common Random Matrix Models

- Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

- Motivation: EEG-based clustering*

- Covariance Distance Inference

- Revisiting Motivation*

- Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

- Support Vector Machines

- Semi-Supervised Learning

- From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Basics of Random Matrix Theory (**Romain COUILLET**)

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models
- Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

- Motivation: EEG-based clustering*
- Covariance Distance Inference
- Revisiting Motivation*
- Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

- Support Vector Machines
- Semi-Supervised Learning
- From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

**Baseline scenario:**  $x_1, \dots, x_n \in \mathbb{R}^p$  (or  $\mathbb{C}^p$ ) i.i.d. with  $E[x_1] = 0$ ,  $E[x_1 x_1^\top] = C_p$ :

**Baseline scenario:**  $x_1, \dots, x_n \in \mathbb{R}^p$  (or  $\mathbb{C}^p$ ) i.i.d. with  $E[x_1] = 0$ ,  $E[x_1 x_1^\top] = C_p$ :

- ▶ If  $x_1 \sim \mathcal{N}(0, C_p)$ , ML estimator for  $C_p$  is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top.$$

## Context

**Baseline scenario:**  $x_1, \dots, x_n \in \mathbb{R}^p$  (or  $\mathbb{C}^p$ ) i.i.d. with  $E[x_1] = 0$ ,  $E[x_1 x_1^\top] = C_p$ :

- ▶ If  $x_1 \sim \mathcal{N}(0, C_p)$ , ML estimator for  $C_p$  is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top.$$

- ▶ If  $n \rightarrow \infty$ , then, **strong law of large numbers**

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

or equivalently, **in spectral norm**

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

## Context

**Baseline scenario:**  $x_1, \dots, x_n \in \mathbb{R}^p$  (or  $\mathbb{C}^p$ ) i.i.d. with  $E[x_1] = 0$ ,  $E[x_1 x_1^\top] = C_p$ :

- ▶ If  $x_1 \sim \mathcal{N}(0, C_p)$ , ML estimator for  $C_p$  is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top.$$

- ▶ If  $n \rightarrow \infty$ , then, **strong law of large numbers**

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

or equivalently, **in spectral norm**

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

## Random Matrix Regime

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

$$\|\hat{C}_p - C_p\| \not\rightarrow 0.$$



## Context

**Baseline scenario:**  $x_1, \dots, x_n \in \mathbb{R}^p$  (or  $\mathbb{C}^p$ ) i.i.d. with  $E[x_1] = 0$ ,  $E[x_1 x_1^\top] = C_p$ :

- ▶ If  $x_1 \sim \mathcal{N}(0, C_p)$ , ML estimator for  $C_p$  is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top.$$

- ▶ If  $n \rightarrow \infty$ , then, **strong law of large numbers**

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

or equivalently, **in spectral norm**

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

## Random Matrix Regime

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

$$\|\hat{C}_p - C_p\| \not\rightarrow 0.$$

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

## Context

**Baseline scenario:**  $x_1, \dots, x_n \in \mathbb{R}^p$  (or  $\mathbb{C}^p$ ) i.i.d. with  $E[x_1] = 0$ ,  $E[x_1 x_1^\top] = C_p$ :

- ▶ If  $x_1 \sim \mathcal{N}(0, C_p)$ , ML estimator for  $C_p$  is the sample covariance matrix (SCM)

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top.$$

- ▶ If  $n \rightarrow \infty$ , then, **strong law of large numbers**

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

or equivalently, **in spectral norm**

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

## Random Matrix Regime

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

$$\|\hat{C}_p - C_p\| \not\rightarrow 0.$$

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

# The Large Dimensional Fallacies

**Setting:**  $x_i \in \mathbb{R}^p$  i.i.d.,  $x_1 \sim \mathcal{CN}(0, I_p)$

# The Large Dimensional Fallacies

**Setting:**  $x_i \in \mathbb{R}^p$  i.i.d.,  $x_1 \sim \mathcal{CN}(0, I_p)$

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

# The Large Dimensional Fallacies

**Setting:**  $x_i \in \mathbb{R}^p$  i.i.d.,  $x_1 \sim \mathcal{CN}(0, I_p)$

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

$$\max_{1 \leq i, j \leq p} \left| [\hat{C}_p - I_p]_{ij} \right| = \max_{1 \leq i, j \leq p} \left| \frac{1}{n} X_{j, \cdot} X_{i, \cdot}^\top - \delta_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

# The Large Dimensional Fallacies

**Setting:**  $x_i \in \mathbb{R}^p$  i.i.d.,  $x_1 \sim \mathcal{CN}(0, I_p)$

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

$$\max_{1 \leq i, j \leq p} \left| [\hat{C}_p - I_p]_{ij} \right| = \max_{1 \leq i, j \leq p} \left| \frac{1}{n} X_{j, \cdot} X_{i, \cdot}^\top - \delta_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

- ▶ however, **eigenvalue mismatch**

$$\begin{aligned} 0 &= \lambda_1(\hat{C}_p) = \dots = \lambda_{p-n}(\hat{C}_p) \leq \lambda_{p-n+1}(\hat{C}_p) \leq \dots \leq \lambda_p(\hat{C}_p) \\ 1 &= \lambda_1(I_p) = \dots = \lambda_{p-n}(I_p) = \lambda_{p-n+1}(\hat{C}_p) = \dots = \lambda_p(I_p) \end{aligned}$$

# The Large Dimensional Fallacies

**Setting:**  $x_i \in \mathbb{R}^p$  i.i.d.,  $x_1 \sim \mathcal{CN}(0, I_p)$

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

$$\max_{1 \leq i, j \leq p} \left| [\hat{C}_p - I_p]_{ij} \right| = \max_{1 \leq i, j \leq p} \left| \frac{1}{n} X_{j, \cdot} X_{i, \cdot}^\top - \delta_{ij} \right| \xrightarrow{\text{a.s.}} 0.$$

- ▶ however, **eigenvalue mismatch**

$$\begin{aligned} 0 &= \lambda_1(\hat{C}_p) = \dots = \lambda_{p-n}(\hat{C}_p) \leq \lambda_{p-n+1}(\hat{C}_p) \leq \dots \leq \lambda_p(\hat{C}_p) \\ 1 &= \lambda_1(I_p) = \dots = \lambda_{p-n}(I_p) = \lambda_{p-n+1}(\hat{C}_p) = \dots = \lambda_p(I_p) \end{aligned}$$

$\Rightarrow$  **no convergence in spectral norm.**

## The Marčenko–Pastur law

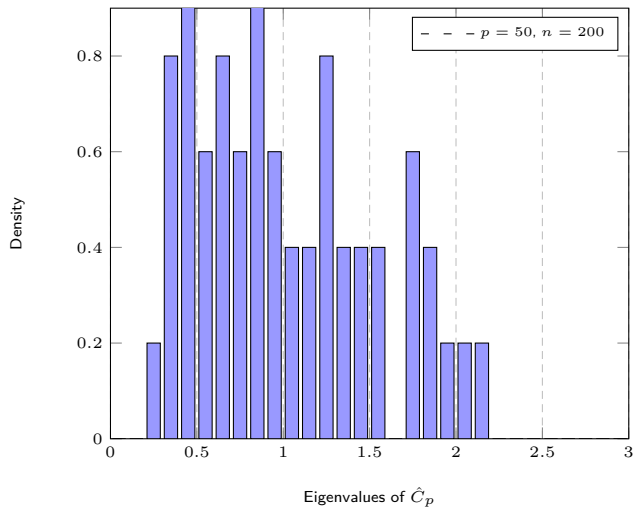


Figure: Histogram of the eigenvalues of  $\hat{C}_p$  for  $c = 1/4$ ,  $C_p = I_p$ .



## The Marčenko–Pastur law

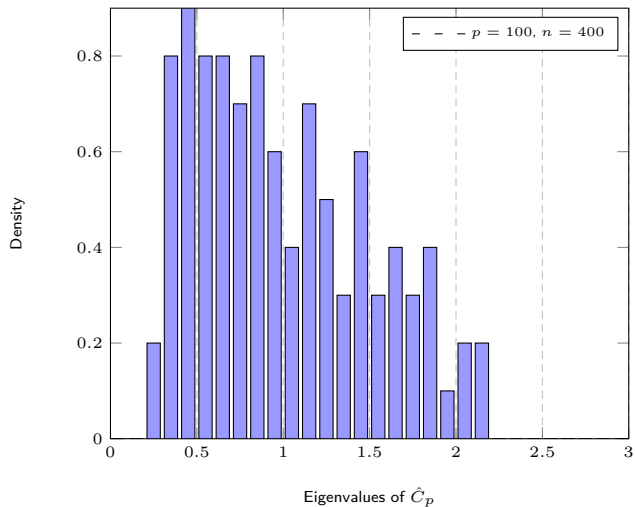


Figure: Histogram of the eigenvalues of  $\hat{C}_p$  for  $c = 1/4$ ,  $C_p = I_p$ .

## The Marčenko–Pastur law

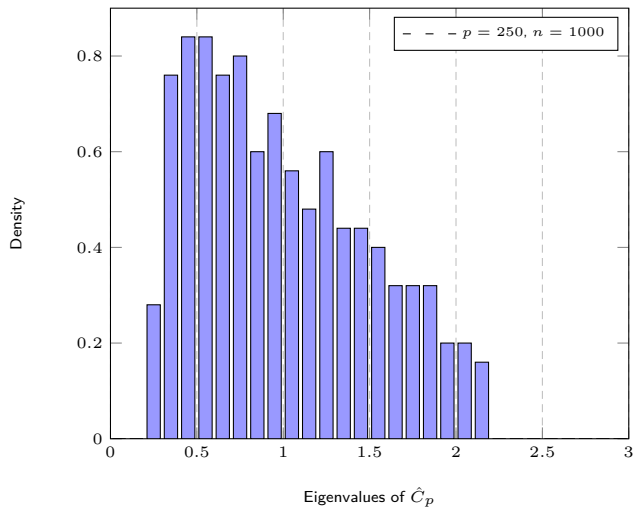


Figure: Histogram of the eigenvalues of  $\hat{C}_p$  for  $c = 1/4, C_p = I_p$ .

## The Marčenko–Pastur law

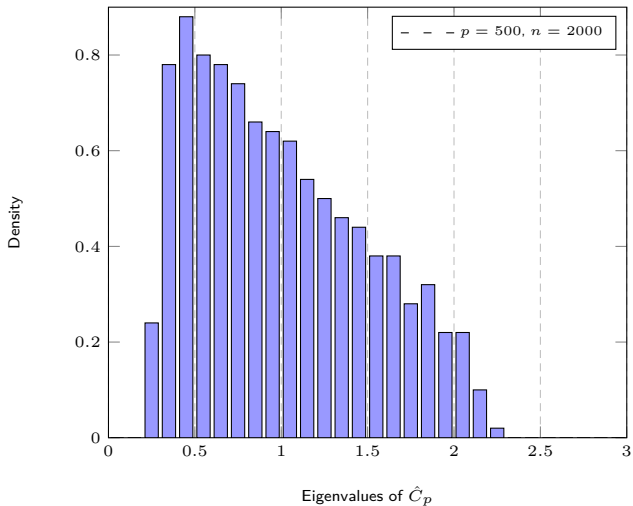


Figure: Histogram of the eigenvalues of  $\hat{C}_p$  for  $c = 1/4$ ,  $C_p = I_p$ .

## The Marčenko–Pastur law

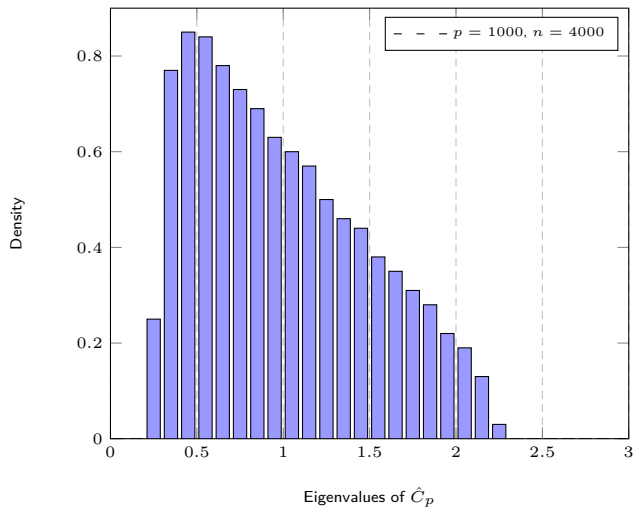


Figure: Histogram of the eigenvalues of  $\hat{C}_p$  for  $c = 1/4, C_p = I_p$ .

## The Marčenko–Pastur law

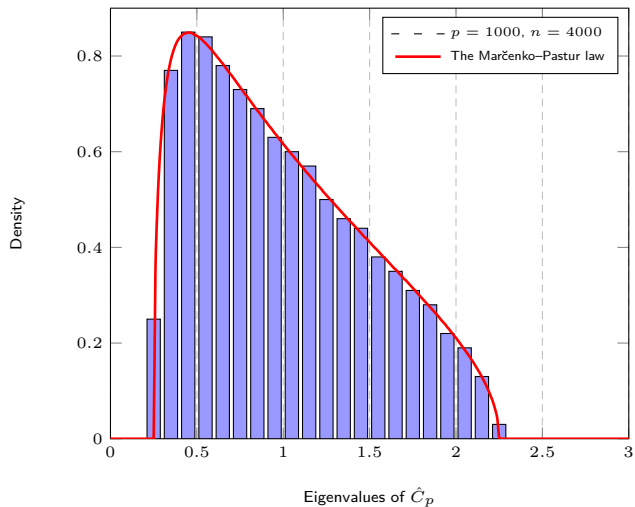


Figure: Histogram of the eigenvalues of  $\hat{C}_p$  for  $c = 1/4$ ,  $C_p = I_p$ .

## Definition (Empirical Spectral Distribution)

Empirical spectral distribution (e.s.d.)  $\mu_p$  of Hermitian matrix  $A_p \in \mathbb{R}^{p \times p}$  is

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

# The Marčenko–Pastur law

## Definition (Empirical Spectral Distribution)

Empirical spectral distribution (e.s.d.)  $\mu_p$  of Hermitian matrix  $A_p \in \mathbb{R}^{p \times p}$  is

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

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

$X_p \in \mathbb{R}^{p \times n}$  with i.i.d. zero mean, unit variance entries.

As  $p, n \rightarrow \infty$  with  $p/n \rightarrow c \in (0, \infty)$ , e.s.d.  $\mu_p$  of  $\frac{1}{n} X_p X_p^T$  satisfies

$$\mu_p \xrightarrow{\text{a.s.}} \mu_{(c)}$$

in distribution (i.e.,  $\int f(t) \mu_p(dt) \xrightarrow{\text{a.s.}} \int f(t) \mu_{(c)}(dt)$  for all bounded continuous  $f$ ), where

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

# The Marčenko–Pastur law

## Definition (Empirical Spectral Distribution)

Empirical spectral distribution (e.s.d.)  $\mu_p$  of Hermitian matrix  $A_p \in \mathbb{R}^{p \times p}$  is

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

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

$X_p \in \mathbb{R}^{p \times n}$  with i.i.d. zero mean, unit variance entries.

As  $p, n \rightarrow \infty$  with  $p/n \rightarrow c \in (0, \infty)$ , e.s.d.  $\mu_p$  of  $\frac{1}{n} X_p X_p^T$  satisfies

$$\mu_p \xrightarrow{\text{a.s.}} \mu_{(c)}$$

in distribution (i.e.,  $\int f(t) \mu_p(dt) \xrightarrow{\text{a.s.}} \int f(t) \mu_{(c)}(dt)$  for all bounded continuous  $f$ ), where

- ▶  $\mu_{(c)}(\{0\}) = \max\{0, 1 - c^{-1}\}$
- ▶ on  $(0, \infty)$ ,  $\mu_{(c)}$  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

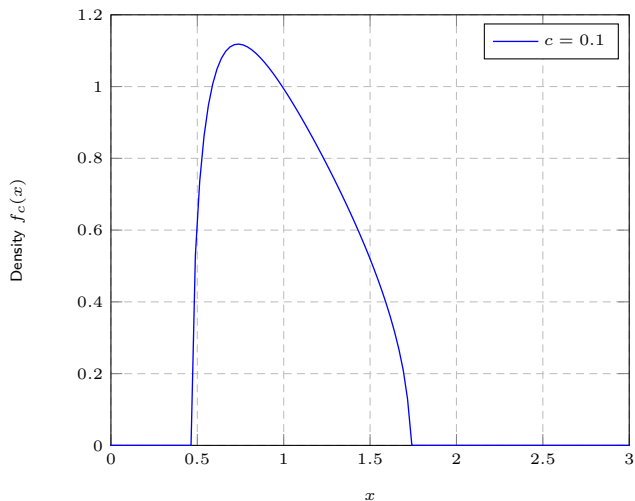


Figure: Marčenko–Pastur law for different limit ratios  $c = \lim_{p \rightarrow \infty} p/n$ .

## The Marčenko–Pastur law

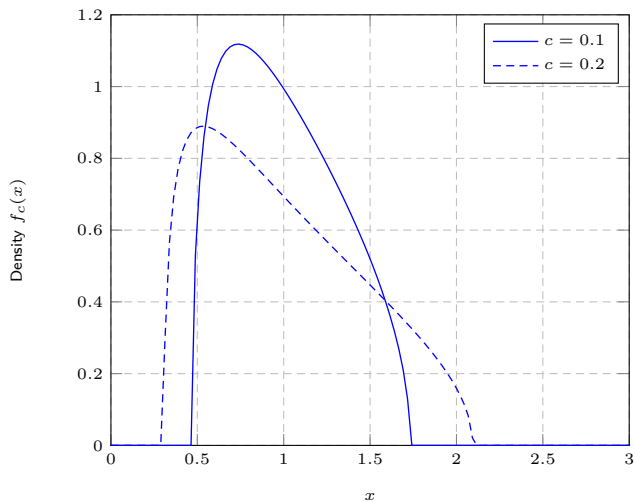


Figure: Marčenko–Pastur law for different limit ratios  $c = \lim_{p \rightarrow \infty} p/n$ .

## The Marčenko–Pastur law

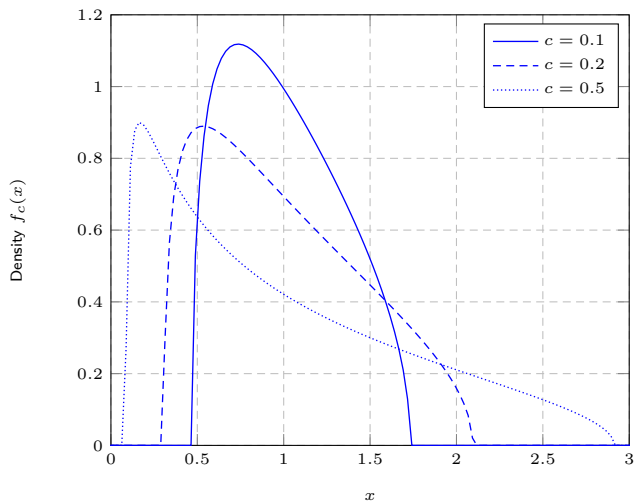


Figure: Marčenko–Pastur law for different limit ratios  $c = \lim_{p \rightarrow \infty} p/n$ .

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

# The Stieltjes transform

## Definition (Stieltjes Transform)

For  $\mu$  real probability measure of support  $\text{supp}(\mu)$ , Stieltjes transform  $m_\mu$  defined, for  $z \in \mathbb{C} \setminus \text{supp}(\mu)$ , as

$$m_\mu(z) = \int \frac{1}{t - z} \mu(dt).$$

# The Stieltjes transform

## Definition (Stieltjes Transform)

For  $\mu$  real probability measure of support  $\text{supp}(\mu)$ , Stieltjes transform  $m_\mu$  defined, for  $z \in \mathbb{C} \setminus \text{supp}(\mu)$ , as

$$m_\mu(z) = \int \frac{1}{t - z} \mu(dt).$$

## Property (Inverse Stieltjes Transform)

For  $a < b$  continuity points of  $\mu$ ,

$$\mu([a, b]) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \int_a^b \Im[m_\mu(x + i\varepsilon)] dx$$

# The Stieltjes transform

## Definition (Stieltjes Transform)

For  $\mu$  real probability measure of support  $\text{supp}(\mu)$ , Stieltjes transform  $m_\mu$  defined, for  $z \in \mathbb{C} \setminus \text{supp}(\mu)$ , as

$$m_\mu(z) = \int \frac{1}{t - z} \mu(dt).$$

## Property (Inverse Stieltjes Transform)

For  $a < b$  continuity points of  $\mu$ ,

$$\mu([a, b]) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \int_a^b \Im[m_\mu(x + i\varepsilon)] dx$$

Besides, if  $\mu$  has a density  $f$  at  $x$ ,

$$f(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \Im[m_\mu(x + i\varepsilon)].$$

# The Stieltjes transform

## Property (Relation to e.s.d.)

If  $\mu$  e.s.d. of Hermitian  $A \in \mathbb{R}^{p \times p}$ , (i.e.,  $\mu = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(A)}$ )

$$m_\mu(z) = \frac{1}{p} \operatorname{tr} (A - zI_p)^{-1}$$



# The Stieltjes transform

## Property (Relation to e.s.d.)

If  $\mu$  e.s.d. of Hermitian  $A \in \mathbb{R}^{p \times p}$ , (i.e.,  $\mu = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(A)}$ )

$$m_\mu(z) = \frac{1}{p} \operatorname{tr} (A - zI_p)^{-1}$$

**Proof:**

$$\begin{aligned} m_\mu(z) &= \int \frac{\mu(dt)}{t-z} = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i(A) - z} = \frac{1}{p} \operatorname{tr} (\operatorname{diag}\{\lambda_i(A)\} - zI_p)^{-1} \\ &= \frac{1}{p} \operatorname{tr} (A - zI_p)^{-1}. \end{aligned}$$

# The Stieltjes transform

## Property (Relation to e.s.d.)

If  $\mu$  e.s.d. of Hermitian  $A \in \mathbb{R}^{p \times p}$ , (i.e.,  $\mu = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(A)}$ )

$$m_\mu(z) = \frac{1}{p} \operatorname{tr} (A - zI_p)^{-1}$$

**Proof:**

$$\begin{aligned} m_\mu(z) &= \int \frac{\mu(dt)}{t - z} = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i(A) - z} = \frac{1}{p} \operatorname{tr} (\operatorname{diag}\{\lambda_i(A)\} - zI_p)^{-1} \\ &= \frac{1}{p} \operatorname{tr} (A - zI_p)^{-1}. \end{aligned}$$

**Fundamental object:** the resolvent of  $A$

$$Q_A(z) \equiv (A - zI_p)^{-1}.$$

# The Stieltjes transform

## Property (Stieltjes transform of Gram matrices)

For  $X \in \mathbb{C}^{p \times n}$ , and

- ▶  $\mu$  e.s.d. of  $XX^T$
- ▶  $\tilde{\mu}$  e.s.d. of  $X^T X$

Then

$$m_{\mu}(z) = \frac{n}{p} m_{\tilde{\mu}}(z) - \frac{p-n}{p} \frac{1}{z}.$$

# The Stieltjes transform

## Property (Stieltjes transform of Gram matrices)

For  $X \in \mathbb{C}^{p \times n}$ , and

- ▶  $\mu$  e.s.d. of  $XX^\top$
- ▶  $\tilde{\mu}$  e.s.d. of  $X^\top X$

Then

$$m_\mu(z) = \frac{n}{p} m_{\tilde{\mu}}(z) - \frac{p-n}{p} \frac{1}{z}.$$

**Proof:**

$$m_\mu(z) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i(XX^\top) - z} = \frac{1}{p} \sum_{i=1}^n \frac{1}{\lambda_i(X^\top X) - z} + \frac{1}{p} (p-n) \frac{1}{0-z}.$$

**Three fundamental lemmas in all proofs.**

## Lemma (Resolvent Identity)

For  $A, B \in \mathbb{R}^{p \times p}$  invertible,

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}.$$

**Proof:** Simply left-multiply by  $A$  and right-multiply by  $B$  on both sides.

# The Stieltjes transform

**Three fundamental lemmas in all proofs.**

## Lemma (Resolvent Identity)

For  $A, B \in \mathbb{R}^{p \times p}$  invertible,

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}.$$

**Proof:** Simply left-multiply by  $A$  and right-multiply by  $B$  on both sides.

## Corollary

For  $t \in \mathbb{C}$ ,  $x \in \mathbb{R}^p$ ,  $A \in \mathbb{R}^{p \times p}$ , with  $A$  and  $A + txx^T$  invertible,

$$(A + txx^T)^{-1}x = \frac{A^{-1}x}{1 + tx^T A^{-1}x}.$$

**Proof Intuition:** Left-multiply by  $(A + tcc^T)$  on both sides.

# The Stieltjes transform

Three fundamental lemmas in all proofs.

## Lemma (Rank-one perturbation)

For  $A, B \in \mathbb{R}^{p \times p}$  Hermitian nonnegative definite, e.s.d.  $\mu$  of  $A$ ,  $t > 0$ ,  $x \in \mathbb{R}^p$ ,  $z \in \mathbb{C} \setminus \text{supp}(\mu)$ ,

$$\left| \frac{1}{p} \text{tr} B (A + txx^T - zI_p)^{-1} - \frac{1}{p} \text{tr} B (A - zI_p)^{-1} \right| \leq \frac{1}{p} \frac{\|B\|}{\text{dist}(z, \text{supp}(\mu))}$$

# The Stieltjes transform

Three fundamental lemmas in all proofs.

## Lemma (Rank-one perturbation)

For  $A, B \in \mathbb{R}^{p \times p}$  Hermitian nonnegative definite, e.s.d.  $\mu$  of  $A$ ,  $t > 0$ ,  $x \in \mathbb{R}^p$ ,  $z \in \mathbb{C} \setminus \text{supp}(\mu)$ ,

$$\left| \frac{1}{p} \text{tr} B (A + txx^T - zI_p)^{-1} - \frac{1}{p} \text{tr} B (A - zI_p)^{-1} \right| \leq \frac{1}{p} \frac{\|B\|}{\text{dist}(z, \text{supp}(\mu))}$$

In particular, as  $p \rightarrow \infty$ , if  $\limsup_p \|B\| < \infty$ ,

$$\frac{1}{p} \text{tr} B (A + txx^T - zI_p)^{-1} - \frac{1}{p} \text{tr} B (A - zI_p)^{-1} \rightarrow 0.$$



# The Stieltjes transform

Three fundamental lemmas in all proofs.

## Lemma (Rank-one perturbation)

For  $A, B \in \mathbb{R}^{p \times p}$  Hermitian nonnegative definite, e.s.d.  $\mu$  of  $A$ ,  $t > 0$ ,  $x \in \mathbb{R}^p$ ,  $z \in \mathbb{C} \setminus \text{supp}(\mu)$ ,

$$\left| \frac{1}{p} \text{tr} B (A + txx^T - zI_p)^{-1} - \frac{1}{p} \text{tr} B (A - zI_p)^{-1} \right| \leq \frac{1}{p} \frac{\|B\|}{\text{dist}(z, \text{supp}(\mu))}$$

In particular, as  $p \rightarrow \infty$ , if  $\limsup_p \|B\| < \infty$ ,

$$\frac{1}{p} \text{tr} B (A + txx^T - zI_p)^{-1} - \frac{1}{p} \text{tr} B (A - zI_p)^{-1} \rightarrow 0.$$

**Proof Intuition:** Based on Weyl's interlacing identity (eigenvalues of  $A$  and  $A + txx^T$  are interlaced).

**Three fundamental lemmas in all proofs.**

## Lemma (Trace Lemma)

For

- ▶  $x \in \mathbb{R}^p$  with i.i.d. entries with zero mean, unit variance, finite  $2k$  order moment,
- ▶  $A \in \mathbb{R}^{p \times p}$  deterministic (or independent of  $x$ ),

then

$$E \left[ \left| \frac{1}{p} x^\top A x - \frac{1}{p} \text{tr} A \right|^k \right] \leq K \frac{\|A\|^p}{p^{k/2}}.$$

# The Stieltjes transform

**Three fundamental lemmas in all proofs.**

## Lemma (Trace Lemma)

For

- ▶  $x \in \mathbb{R}^p$  with i.i.d. entries with zero mean, unit variance, finite  $2k$  order moment,
- ▶  $A \in \mathbb{R}^{p \times p}$  deterministic (or independent of  $x$ ),

then

$$E \left[ \left| \frac{1}{p} x^\top A x - \frac{1}{p} \operatorname{tr} A \right|^k \right] \leq K \frac{\|A\|^p}{p^{k/2}}.$$

In particular, if  $\limsup_p \|A\| < \infty$ , and  $x$  has entries with finite eighth-order moment,

$$\frac{1}{p} x^\top A x - \frac{1}{p} \operatorname{tr} A \xrightarrow{\text{a.s.}} 0$$

(by Markov inequality and Borel Cantelli lemma).

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

$X_p \in \mathbb{R}^{p \times n}$  with i.i.d. zero mean, unit variance entries.

As  $p, n \rightarrow \infty$  with  $p/n \rightarrow c \in (0, \infty)$ , e.s.d.  $\mu_p$  of  $\frac{1}{n} X_p X_p^T$  satisfies

$$\mu_p \xrightarrow{\text{a.s.}} \mu_{(c)}$$

weakly, where

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

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

$X_p \in \mathbb{R}^{p \times n}$  with i.i.d. zero mean, unit variance entries.

As  $p, n \rightarrow \infty$  with  $p/n \rightarrow c \in (0, \infty)$ , e.s.d.  $\mu_p$  of  $\frac{1}{n} X_p X_p^T$  satisfies

$$\mu_p \xrightarrow{\text{a.s.}} \mu_{(c)}$$

weakly, where

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

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

**Stieltjes transform approach.**

# Proof of the Marčenko–Pastur law

**Stieltjes transform approach.**

Proof

► With  $\mu_p$  e.s.d. of  $\frac{1}{n}X_pX_p^\top$ ,

$$m_{\mu_p}(z) = \frac{1}{p} \operatorname{tr} \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} = \frac{1}{p} \sum_{i=1}^p \left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{ii}.$$

# Proof of the Marčenko–Pastur law

Stieltjes transform approach.

Proof

- ▶ With  $\mu_p$  e.s.d. of  $\frac{1}{n}X_pX_p^\top$ ,

$$m_{\mu_p}(z) = \frac{1}{p} \operatorname{tr} \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} = \frac{1}{p} \sum_{i=1}^p \left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{ii}.$$

- ▶ Write

$$X_p = \begin{bmatrix} y^\top \\ Y_{p-1} \end{bmatrix} \in \mathbb{R}^{p \times n}$$



# Proof of the Marčenko–Pastur law

Stieltjes transform approach.

Proof

- ▶ With  $\mu_p$  e.s.d. of  $\frac{1}{n}X_pX_p^\top$ ,

$$m_{\mu_p}(z) = \frac{1}{p} \operatorname{tr} \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} = \frac{1}{p} \sum_{i=1}^p \left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{ii}.$$

- ▶ Write

$$X_p = \begin{bmatrix} y^\top \\ Y_{p-1} \end{bmatrix} \in \mathbb{R}^{p \times n}$$

so that, for  $\Im[z] > 0$ ,

$$\left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} = \begin{pmatrix} \frac{1}{n} y^\top y - z & \frac{1}{n} y^\top Y_{p-1} \\ \frac{1}{n} Y_{p-1} y & \frac{1}{n} Y_{p-1} Y_{p-1}^\top - z I_{p-1} \end{pmatrix}^{-1}.$$

## Proof (continued)

- ▶ From block matrix inverse formula

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(A - BD^{-1}C)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

we have

$$\left[ \left( \frac{1}{n} X_p X_p^T - z I_p \right)^{-1} \right]_{11} = \frac{1}{-z - z \frac{1}{n} \mathbf{y}^T \left( \frac{1}{n} Y_{p-1}^T Y_{p-1} - z I_n \right)^{-1} \mathbf{y}}.$$

## Proof (continued)

- ▶ From block matrix inverse formula

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(A - BD^{-1}C)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

we have

$$\left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} = \frac{1}{-z - z \frac{1}{n} \mathbf{y}^\top \left( \frac{1}{n} Y_{p-1}^\top Y_{p-1} - z I_n \right)^{-1} \mathbf{y}}.$$

- ▶ By **Trace Lemma**, as  $p, n \rightarrow \infty$

$$\left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} - \frac{1}{-z - z \frac{1}{n} \text{tr} \left( \frac{1}{n} Y_{p-1}^\top Y_{p-1} - z I_n \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

## Proof of the Marčenko–Pastur law

### Proof (continued)

- ▶ By **Rank-1 Perturbation Lemma** ( $X_p^\top X_p = Y_{p-1}^\top Y_{p-1} + yy^\top$ ), as  $p, n \rightarrow \infty$

$$\left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} = \frac{1}{-z - z \frac{1}{n} \text{tr} \left( \frac{1}{n} X_p^\top X_p - z I_n \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

# Proof of the Marčenko–Pastur law

## Proof (continued)

- ▶ By **Rank-1 Perturbation Lemma** ( $X_p^\top X_p = Y_{p-1}^\top Y_{p-1} + yy^\top$ ), as  $p, n \rightarrow \infty$

$$\left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} = \frac{1}{-z - z \frac{1}{n} \operatorname{tr} \left( \frac{1}{n} X_p^\top X_p - z I_n \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

- ▶ Since  $\frac{1}{n} \operatorname{tr} \left( \frac{1}{n} X_p^\top X_p - z I_n \right)^{-1} = \frac{1}{n} \operatorname{tr} \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} - \frac{n-p}{n} \frac{1}{z}$ ,

$$\left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} = \frac{1}{1 - \frac{p}{n} - z - z \frac{1}{n} \operatorname{tr} \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

# Proof of the Marčenko–Pastur law

## Proof (continued)

- ▶ By **Rank-1 Perturbation Lemma** ( $X_p^\top X_p = Y_{p-1}^\top Y_{p-1} + yy^\top$ ), as  $p, n \rightarrow \infty$

$$\left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} - \frac{1}{-z - z \frac{1}{n} \operatorname{tr} \left( \frac{1}{n} X_p^\top X_p - z I_n \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

- ▶ Since  $\frac{1}{n} \operatorname{tr} \left( \frac{1}{n} X_p^\top X_p - z I_n \right)^{-1} = \frac{1}{n} \operatorname{tr} \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} - \frac{n-p}{n} \frac{1}{z}$ ,

$$\left[ \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} \right]_{11} - \frac{1}{1 - \frac{p}{n} - z - z \frac{1}{n} \operatorname{tr} \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1}} \xrightarrow{\text{a.s.}} 0.$$

- ▶ Repeating for **entries**  $(2, 2), \dots, (p, p)$ , and averaging, we get (for  $\Im[z] > 0$ )

$$m_{\mu_p}(z) - \frac{1}{1 - \frac{p}{n} - z - z \frac{p}{n} m_{\mu_p}(z)} \xrightarrow{\text{a.s.}} 0.$$

## Proof of the Marčenko–Pastur law

### Proof (continued)

► Then  $m_{\mu_p}(z) \xrightarrow{\text{a.s.}} m(z)$  solution to

$$m(z) = \frac{1}{1 - c - z - czm(z)}$$

## Proof of the Marčenko–Pastur law

### Proof (continued)

► Then  $m_{\mu_p}(z) \xrightarrow{\text{a.s.}} m(z)$  solution to

$$m(z) = \frac{1}{1 - c - z - czm(z)}$$

i.e., (with branch of  $\sqrt{f(z)}$  such that  $m(z) \rightarrow 0$  as  $|z| \rightarrow \infty$ )

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



## Proof of the Marčenko–Pastur law

### Proof (continued)

- Then  $m_{\mu_p}(z) \xrightarrow{\text{a.s.}} m(z)$  solution to

$$m(z) = \frac{1}{1 - c - z - czm(z)}$$

i.e., (with branch of  $\sqrt{f(z)}$  such that  $m(z) \rightarrow 0$  as  $|z| \rightarrow \infty$ )

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

- Finally, by **inverse Stieltjes Transform**, for  $x > 0$ ,

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \Im[m(x + i\varepsilon)] = \frac{\sqrt{((1 + \sqrt{c})^2 - x)(x - (1 - \sqrt{c})^2)}}{2\pi cx} \mathbf{1}_{\{x \in [(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]\}}.$$

And for  $x = 0$ ,

$$\lim_{\varepsilon \downarrow 0} i\varepsilon \Im[m(i\varepsilon)] = (1 - c^{-1}) \mathbf{1}_{\{c > 1\}}.$$

## Sample Covariance Matrices

### Theorem (Sample Covariance Matrix Model [Silverstein, Bai'95])

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{C}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,
- ▶  $X_p \in \mathbb{C}^{p \times n}$  has i.i.d. entries of zero mean and unit variance.

As  $p, n \rightarrow \infty$ ,  $p/n \rightarrow c \in (0, \infty)$ ,  $\tilde{\mu}_p$  e.s.d. of  $\frac{1}{n} Y_p^T Y_p \in \mathbb{R}^{n \times n}$  satisfies

$$\tilde{\mu}_p \xrightarrow{\text{a.s.}} \tilde{\mu}$$

weakly, with  $m_{\tilde{\mu}}(z)$ ,  $\Im[z] > 0$ , unique solution with  $\Im[m_{\tilde{\mu}}(z)] > 0$  of

$$m_{\tilde{\mu}}(z) = \left( -z + c \int \frac{t}{1 + tm_{\tilde{\mu}}(z)} \nu(dt) \right)^{-1}.$$

## Sample Covariance Matrices

### Theorem (Sample Covariance Matrix Model [Silverstein, Bai'95])

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{C}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,
- ▶  $X_p \in \mathbb{C}^{p \times n}$  has i.i.d. entries of zero mean and unit variance.

As  $p, n \rightarrow \infty$ ,  $p/n \rightarrow c \in (0, \infty)$ ,  $\tilde{\mu}_p$  e.s.d. of  $\frac{1}{n} Y_p^T Y_p \in \mathbb{R}^{n \times n}$  satisfies

$$\tilde{\mu}_p \xrightarrow{\text{a.s.}} \tilde{\mu}$$

weakly, with  $m_{\tilde{\mu}}(z)$ ,  $\Im[z] > 0$ , unique solution with  $\Im[m_{\tilde{\mu}}(z)] > 0$  of

$$m_{\tilde{\mu}}(z) = \left( -z + c \int \frac{t}{1 + tm_{\tilde{\mu}}(z)} \nu(dt) \right)^{-1}.$$

Moreover,  $\tilde{\mu}$  is continuous on  $\mathbb{R}^+$  and real analytic wherever positive.

## Sample Covariance Matrices

### Theorem (Sample Covariance Matrix Model [Silverstein, Bai'95])

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{C}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,
- ▶  $X_p \in \mathbb{C}^{p \times n}$  has i.i.d. entries of zero mean and unit variance.

As  $p, n \rightarrow \infty$ ,  $p/n \rightarrow c \in (0, \infty)$ ,  $\tilde{\mu}_p$  e.s.d. of  $\frac{1}{n} Y_p^T Y_p \in \mathbb{R}^{n \times n}$  satisfies

$$\tilde{\mu}_p \xrightarrow{\text{a.s.}} \tilde{\mu}$$

weakly, with  $m_{\tilde{\mu}}(z)$ ,  $\Im[z] > 0$ , unique solution with  $\Im[m_{\tilde{\mu}}(z)] > 0$  of

$$m_{\tilde{\mu}}(z) = \left( -z + c \int \frac{t}{1 + tm_{\tilde{\mu}}(z)} \nu(dt) \right)^{-1}.$$

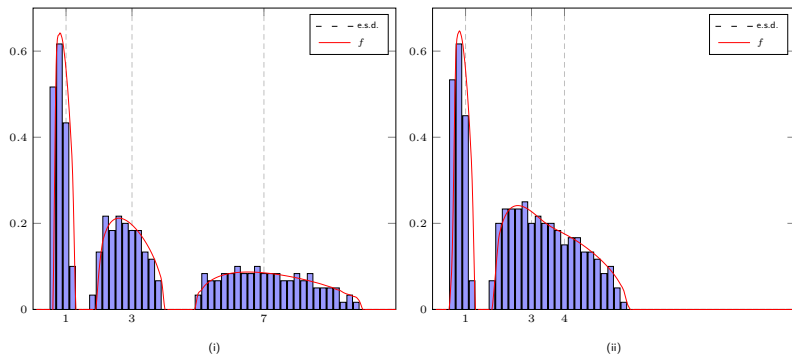
Moreover,  $\tilde{\mu}$  is continuous on  $\mathbb{R}^+$  and real analytic wherever positive.

**Immediate corollary:** For  $\mu_p$  e.s.d. of  $\frac{1}{n} Y_p Y_p^T = \frac{1}{n} \sum_{i=1}^n C_p^{\frac{1}{2}} x_i x_i^T C_p^{\frac{1}{2}}$ ,

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

weakly, with  $\tilde{\mu} = c\mu + (1 - c)\delta_0$ .

## Sample Covariance Matrices



**Figure:** Histogram of the eigenvalues of  $\frac{1}{n} Y_p Y_p^T$ ,  $n = 3000$ ,  $p = 300$ , with  $C_p$  diagonal with evenly weighted masses in (i) 1, 3, 7, (ii) 1, 3, 4.

**Sometimes,  $\mu_p$  does not converge!**

**Sometimes,  $\mu_p$  does not converge!**

- ▶ if  $\nu_p$  does not converge

**Sometimes,  $\mu_p$  does not converge!**

- ▶ if  $\nu_p$  does not converge
- ▶ if  $p/n$  does not converge



**Sometimes,  $\mu_p$  does not converge!**

- ▶ if  $\nu_p$  does not converge
- ▶ if  $p/n$  does not converge
- ▶ if eigenvectors of deterministic matrices play a role!

## Further Models and Deterministic Equivalents

**Sometimes,  $\mu_p$  does not converge!**

- ▶ if  $\nu_p$  does not converge
- ▶ if  $p/n$  does not converge
- ▶ if eigenvectors of deterministic matrices play a role!

**Deterministic equivalents:** sequence  $\bar{\mu}_p$  of **deterministic** measures, with

$$\mu_p - \bar{\mu}_p \xrightarrow{\text{a.s.}} 0$$

## Further Models and Deterministic Equivalents

Sometimes,  $\mu_p$  does not converge!

- ▶ if  $\nu_p$  does not converge
- ▶ if  $p/n$  does not converge
- ▶ if eigenvectors of deterministic matrices play a role!

**Deterministic equivalents:** sequence  $\bar{\mu}_p$  of **deterministic** measures, with

$$\mu_p - \bar{\mu}_p \xrightarrow{\text{a.s.}} 0$$

or equivalently, **deterministic** sequence of  $m_p$  with

$$m_{\mu_p} - m_p \xrightarrow{\text{a.s.}} 0.$$

### Theorem (Doubly-correlated i.i.d. matrices)

Let  $B_p = C_p^{\frac{1}{2}} X_p T_p X_p^T C_p^{\frac{1}{2}}$ , with e.s.d.  $\mu_p$ ,  $X_p \in \mathbb{R}^{p \times n}$  with i.i.d. entries of zero mean, variance  $1/n$ ,  $C_p$  Hermitian nonnegative definite,  $T_p$  diagonal nonnegative,  $\limsup_p \max(\|C_p\|, \|T_p\|) < \infty$ . Denote  $c = p/n$ .

Then, as  $p, n \rightarrow \infty$  with bounded ratio  $c$ , for  $z \in \mathbb{C} \setminus \mathbb{R}^-$ ,

$$m_{\mu_p}(z) - m_p(z) \xrightarrow{\text{a.s.}} 0, \quad m_p(z) = \frac{1}{p} \text{tr} (-zI_p + \bar{e}_p(z)C_p)^{-1}$$

with  $\bar{e}(z)$  unique solution in  $\{z \in \mathbb{C}^+, \bar{e}_p(z) \in \mathbb{C}^+\}$  or  $\{z \in \mathbb{R}^-, \bar{e}_p(z) \in \mathbb{R}^+\}$  of

$$e_p(z) = \frac{1}{p} \text{tr} C_p (-zI_p + \bar{e}_p(z)C_p)^{-1}$$

$$\bar{e}_p(z) = \frac{1}{n} \text{tr} T_p (I_n + ce_p(z)T_p)^{-1}.$$

**Side note on other models.**

Similar results for multiple matrix models:

### Side note on other models.

Similar results for multiple matrix models:

- ▶ **Information-plus-noise:**  $Y_p = A_p + X_p$ ,  $A_p$  deterministic
- ▶ **Variance profile:**  $Y_p = P_p \odot X_p$  (entry-wise product)
- ▶ **Per-column covariance:**  $Y_p = [y_1, \dots, y_n]$ ,  $y_i = C_{p,i}^{\frac{1}{2}} x_i$
- ▶ etc.

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

**Spiked Models**

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support **[Silverstein, Bai'98]**)

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{R}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,



## No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support [**Silverstein, Bai'98**])

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{R}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,
- ▶  $X_p \in \mathbb{R}^{p \times n}$  has i.i.d. entries of zero mean and unit variance,

## No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support [**Silverstein, Bai'98**])

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{R}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,
- ▶  $X_p \in \mathbb{R}^{p \times n}$  has i.i.d. entries of zero mean and unit variance,
- ▶  $E[|X_p|_{ij}^4] < \infty$ ,

## No Eigenvalue Outside the Support

Theorem (No Eigenvalue Outside the Support [**Silverstein, Bai'98**])

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{R}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,
- ▶  $X_p \in \mathbb{R}^{p \times n}$  has i.i.d. entries of zero mean and unit variance,
- ▶  $E[|X_p|_{ij}^4] < \infty$ ,
- ▶  $\max_i \text{dist}(\lambda_i(C_p), \text{supp}(\nu)) \rightarrow 0$ .

## No Eigenvalue Outside the Support

### Theorem (No Eigenvalue Outside the Support [Silverstein, Bai'98])

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{R}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,
- ▶  $X_p \in \mathbb{R}^{p \times n}$  has i.i.d. entries of zero mean and unit variance,
- ▶  $E[|X_p|_{ij}^4] < \infty$ ,
- ▶  $\max_i \text{dist}(\lambda_i(C_p), \text{supp}(\nu)) \rightarrow 0$ .

Let  $\tilde{\mu}$  be the limiting e.s.d. of  $\frac{1}{n} Y_p^T Y_p$  as before. Let  $[a, b] \subset \mathbb{R}^T \setminus \text{supp}(\tilde{\nu})$ . Then,

$$\left\{ \lambda_i \left( \frac{1}{n} Y_p^T Y_p \right) \right\}_{i=1}^n \cap [a, b] = \emptyset$$

for all large  $n$ , almost surely.

## No Eigenvalue Outside the Support

### Theorem (No Eigenvalue Outside the Support [Silverstein, Bai'98])

Let  $Y_p = C_p^{\frac{1}{2}} X_p \in \mathbb{R}^{p \times n}$ , with

- ▶  $C_p \in \mathbb{R}^{p \times p}$  nonnegative definite with e.s.d.  $\nu_p \rightarrow \nu$  weakly,
- ▶  $X_p \in \mathbb{R}^{p \times n}$  has i.i.d. entries of zero mean and unit variance,
- ▶  $E[|X_p|_{ij}^4] < \infty$ ,
- ▶  $\max_i \text{dist}(\lambda_i(C_p), \text{supp}(\nu)) \rightarrow 0$ .

Let  $\tilde{\mu}$  be the limiting e.s.d. of  $\frac{1}{n} Y_p^T Y_p$  as before. Let  $[a, b] \subset \mathbb{R}^T \setminus \text{supp}(\tilde{\nu})$ . Then,

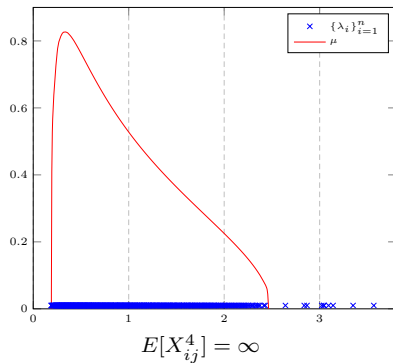
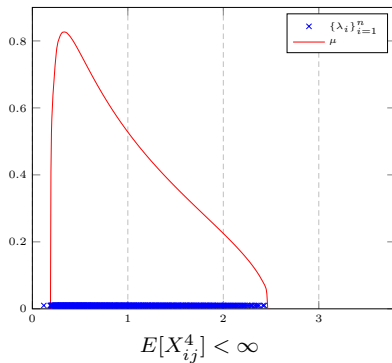
$$\left\{ \lambda_i \left( \frac{1}{n} Y_p^T Y_p \right) \right\}_{i=1}^n \cap [a, b] = \emptyset$$

for all large  $n$ , almost surely.

**In practice:** This means that eigenvalues of  $\frac{1}{n} Y_p^T Y_p$  cannot be bound at macroscopic distance from the bulk, for  $p, n$  large.

Breaking the rules. If we break

- ▶ **Rule 1:** Infinitely many eigenvalues may wander away from  $\text{supp}(\mu)$ .



# Spiked Models

If we break:

- ▶ **Rule 2:**  $C_p$  may create isolated eigenvalues in  $\frac{1}{n} Y_p Y_p^T$ , called spikes.

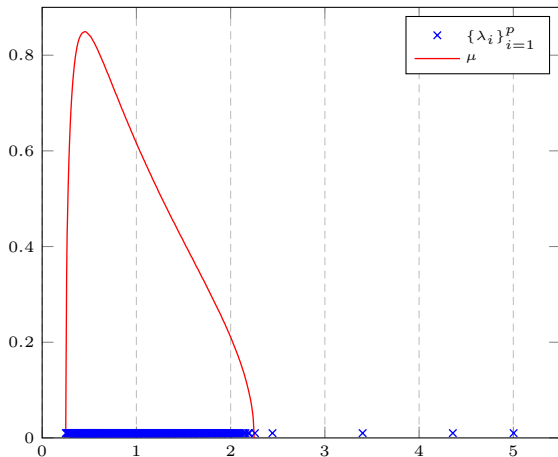


Figure: Eigenvalues of  $\frac{1}{n} Y_p Y_p^T$ ,  $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-4}, 2, 3, 4, 5)$ ,  $p = 500$ ,  $n = 2000$ .

## Spiked Models: The phase transition phenomenon

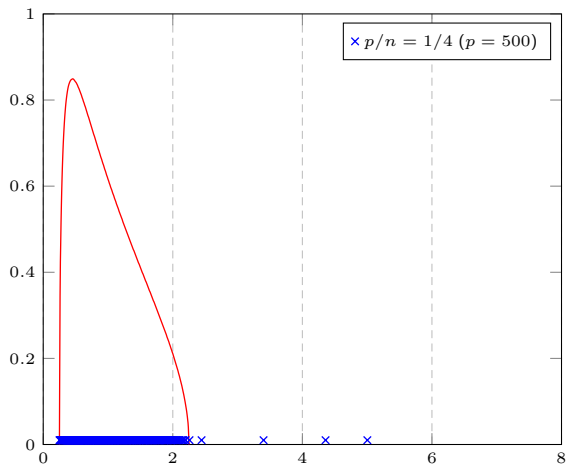


Figure: Eigenvalues of  $\frac{1}{n} Y_p Y_p^T$ ,  $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-4}, 2, 3, 4, 5)$ .



## Spiked Models: The phase transition phenomenon

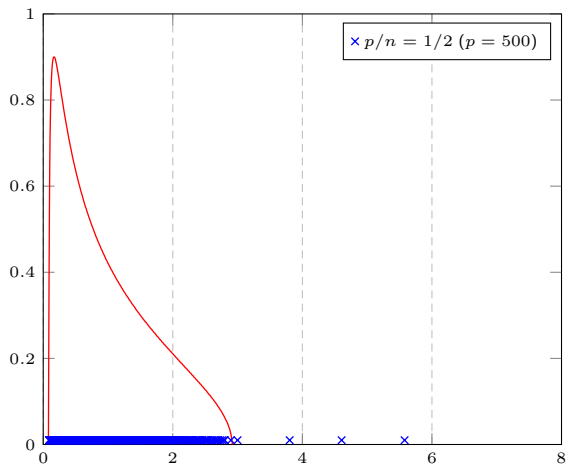


Figure: Eigenvalues of  $\frac{1}{n} Y_p Y_p^T$ ,  $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-4}, 2, 3, 4, 5)$ .

## Spiked Models: The phase transition phenomenon

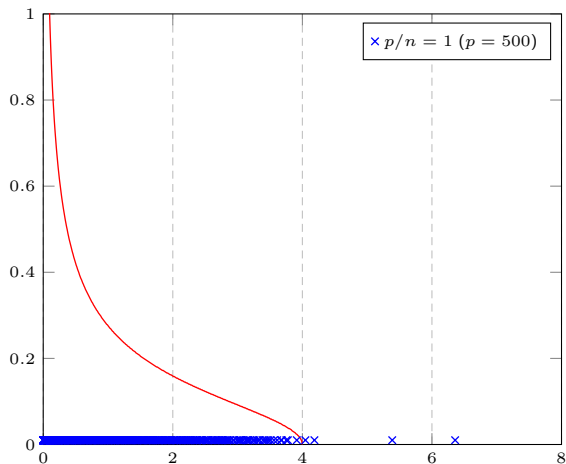


Figure: Eigenvalues of  $\frac{1}{n} Y_p Y_p^T$ ,  $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-4}, 2, 3, 4, 5)$ .

## Spiked Models: The phase transition phenomenon

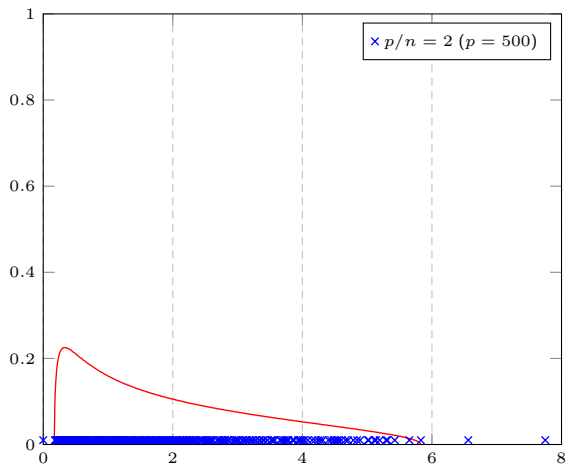


Figure: Eigenvalues of  $\frac{1}{n} Y_p Y_p^T$ ,  $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-4}, 2, 3, 4, 5)$ .

## 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,  $E[|X_p|_{ij}^4] < \infty$ .
- ▶  $C_p = I_p + P$ ,  $P = U\Omega U^T$ , where, for  $K$  fixed,

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

## 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,  $E[|X_p|_{ij}^4] < \infty$ .
- ▶  $C_p = I_p + P$ ,  $P = U\Omega U^T$ , where, for  $K$  fixed,

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

Then, as  $p, n \rightarrow \infty$ ,  $p/n \rightarrow c \in (0, \infty)$ , denoting  $\lambda_i = \lambda_i(\frac{1}{n} Y_p Y_p^T)$ ,

- ▶ if  $\omega_m > \sqrt{c}$ ,

$$\lambda_m \xrightarrow{\text{a.s.}} 1 + \omega_m + c \frac{1 + \omega_m}{\omega_m} > (1 + \sqrt{c})^2$$

## 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,  $E[|X_p|_{ij}^4] < \infty$ .
- ▶  $C_p = I_p + P$ ,  $P = U\Omega U^T$ , where, for  $K$  fixed,

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

Then, as  $p, n \rightarrow \infty$ ,  $p/n \rightarrow c \in (0, \infty)$ , denoting  $\lambda_i = \lambda_i(\frac{1}{n} Y_p Y_p^T)$ ,

- ▶ if  $\omega_m > \sqrt{c}$ ,

$$\lambda_m \xrightarrow{\text{a.s.}} 1 + \omega_m + c \frac{1 + \omega_m}{\omega_m} > (1 + \sqrt{c})^2$$

- ▶ if  $\omega_m \in (0, \sqrt{c}]$ ,

$$\lambda_m \xrightarrow{\text{a.s.}} (1 + \sqrt{c})^2$$

## Spiked Models

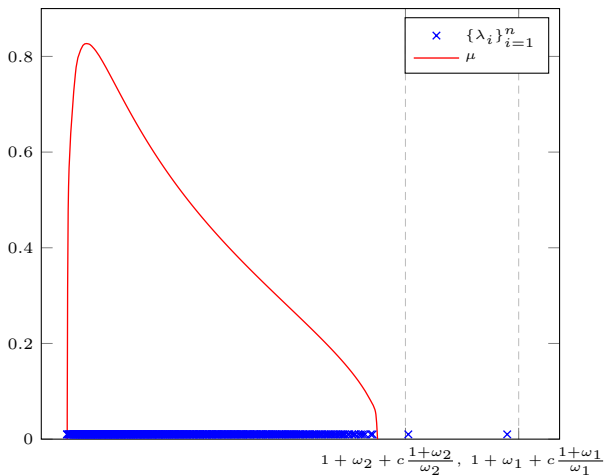


Figure: Eigenvalues of  $\frac{1}{n} Y_p Y_p^T$ ,  $C_p = \text{diag}(\underbrace{1, \dots, 1}_{p-2}, 2, 3)$ ,  $p = 500$ ,  $n = 1500$ .

# Spiked Models

## Proof

- ▶ **Two ingredients:** Algebraic calculus + trace lemma



## Proof

- ▶ **Two ingredients:** Algebraic calculus + trace lemma
- ▶ **Find eigenvalues away from eigenvalues of  $\frac{1}{n}X_pX_p^\top$ :**

$$\begin{aligned}0 &= \det\left(\frac{1}{n}Y_pY_p^\top - \lambda I_p\right), \quad Y_p = C_p^{\frac{1}{2}}X_p \\ &= \det(C_p) \det\left(\frac{1}{n}X_pX_p^\top - \lambda C_p^{-1}\right) \\ &= \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p + \lambda(I_p - C_p^{-1})\right) \\ &= \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right) \det\left(I_p + \lambda(I_p - C_p^{-1})\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right)^{-1}\right).\end{aligned}$$

# Spiked Models

## Proof

- ▶ **Two ingredients:** Algebraic calculus + trace lemma
- ▶ **Find eigenvalues away from eigenvalues of  $\frac{1}{n}X_pX_p^\top$ :**

$$\begin{aligned}0 &= \det\left(\frac{1}{n}Y_pY_p^\top - \lambda I_p\right), \quad Y_p = C_p^{\frac{1}{2}}X_p \\&= \det(C_p) \det\left(\frac{1}{n}X_pX_p^\top - \lambda C_p^{-1}\right) \\&= \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p + \lambda(I_p - C_p^{-1})\right) \\&= \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right) \det\left(I_p + \lambda(I_p - C_p^{-1})\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right)^{-1}\right).\end{aligned}$$

- ▶ **Use low rank property:** ( $C_p = I_p + P = I_p + U\Omega U^\top$ )

$$I_p - C_p^{-1} = I_p - (I_p + U\Omega U^\top)^{-1} = U(I_K + \Omega^{-1})^{-1}U^\top, \quad \Omega \in \mathbb{C}^{K \times K}.$$

Hence

$$0 = \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right) \det\left(I_p + \lambda U(I_K + \Omega^{-1})^{-1}U^\top\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right)^{-1}\right).$$

## Proof

- ▶ **Two ingredients:** Algebraic calculus + trace lemma
- ▶ **Find eigenvalues away from eigenvalues of  $\frac{1}{n}X_pX_p^\top$ :**

$$\begin{aligned}0 &= \det\left(\frac{1}{n}Y_pY_p^\top - \lambda I_p\right), \quad Y_p = C_p^{\frac{1}{2}}X_p \\ &= \det(C_p) \det\left(\frac{1}{n}X_pX_p^\top - \lambda C_p^{-1}\right) \\ &= \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p + \lambda(I_p - C_p^{-1})\right) \\ &= \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right) \det\left(I_p + \lambda(I_p - C_p^{-1})\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right)^{-1}\right).\end{aligned}$$

- ▶ **Use low rank property:** ( $C_p = I_p + P = I_p + U\Omega U^\top$ )

$$I_p - C_p^{-1} = I_p - (I_p + U\Omega U^\top)^{-1} = U(I_K + \Omega^{-1})^{-1}U^\top, \quad \Omega \in \mathbb{C}^{K \times K}.$$

Hence

$$0 = \det\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right) \det\left(I_p + \lambda U(I_K + \Omega^{-1})^{-1}U^\top\left(\frac{1}{n}X_pX_p^\top - \lambda I_p\right)^{-1}\right).$$

## Proof (2)

► **Sylvester's identity** ( $\det(I + AB) = \det(I + BA)$ ),

$$0 = \det\left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right) \det\left(I_K + \lambda(I_K + \Omega^{-1})^{-1}U^\top \left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right)^{-1}U\right)$$

## Proof (2)

- ▶ **Sylvester's identity** ( $\det(I + AB) = \det(I + BA)$ ),

$$0 = \det\left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right) \det\left(I_K + \lambda(I_K + \Omega^{-1})^{-1}U^\top \left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right)^{-1}U\right)$$

- ▶ **No eigenvalue outside the support [Bai,Sil'98]**:  $\det(\frac{1}{n}X_p X_p^\top - \lambda I_p)$  has no zero beyond  $(1 + \sqrt{c})^2$  for all large  $n$  a.s.

## Proof (2)

- ▶ **Sylvester's identity** ( $\det(I + AB) = \det(I + BA)$ ),

$$0 = \det\left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right) \det\left(I_K + \lambda(I_K + \Omega^{-1})^{-1}U^\top\left(\frac{1}{n}X_p X_p^\top - \lambda I_p\right)^{-1}U\right)$$

- ▶ **No eigenvalue outside the support [Bai, Sil'98]**:  $\det(\frac{1}{n}X_p X_p^\top - \lambda I_p)$  has no zero beyond  $(1 + \sqrt{c})^2$  for all large  $n$  a.s.
- ▶ **Extension of Trace Lemma**: for each  $z \in \mathbb{C} \setminus \text{supp}(\mu)$ ,

$$U^\top\left(\frac{1}{n}X_p X_p^\top - z I_p\right)^{-1}U \xrightarrow{\text{a.s.}} m_\mu(z)I_K.$$

( $X_p$  being “almost-unitarily invariant”,  $U$  made of “i.i.d.-like” random vectors)

## Proof (2)

- ▶ **Sylvester's identity** ( $\det(I + AB) = \det(I + BA)$ ),

$$0 = \det \left( \frac{1}{n} X_p X_p^\top - \lambda I_p \right) \det \left( I_K + \lambda (I_K + \Omega^{-1})^{-1} U^\top \left( \frac{1}{n} X_p X_p^\top - \lambda I_p \right)^{-1} U \right)$$

- ▶ **No eigenvalue outside the support [Bai, Sil'98]**:  $\det(\frac{1}{n} X_p X_p^\top - \lambda I_p)$  has no zero beyond  $(1 + \sqrt{c})^2$  for all large  $n$  a.s.
- ▶ **Extension of Trace Lemma**: for each  $z \in \mathbb{C} \setminus \text{supp}(\mu)$ ,

$$U^\top \left( \frac{1}{n} X_p X_p^\top - z I_p \right)^{-1} U \xrightarrow{\text{a.s.}} m_\mu(z) I_K.$$

( $X_p$  being “almost-unitarily invariant”,  $U$  made of “i.i.d.-like” random vectors)

- ▶ As a result, for all large  $n$  a.s.,

$$\begin{aligned} 0 &= \det \left( I_K + \lambda (I_K + \Omega^{-1})^{-1} U^\top \left( \frac{1}{n} X_p X_p^\top - \lambda I_p \right)^{-1} U \right) \\ &\simeq \prod_{k=1}^K \left( 1 + \frac{\lambda}{1 + \omega_k^{-1}} m_\mu(\lambda) \right) = \prod_{k=1}^K \left( 1 + \frac{\omega_k}{1 + \omega_k} \lambda m_\mu(\lambda) \right) \end{aligned}$$

## Proof (3)

- ▶ **Limiting solutions:** zeros of

$$\lambda m_{\mu}(\lambda) = -\frac{1 + \omega_m}{\omega_m}.$$



## Proof (3)

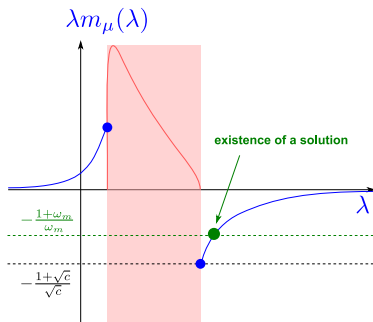
- ▶ **Limiting solutions:** zeros of

$$\lambda m_\mu(\lambda) = -\frac{1 + \omega_m}{\omega_m}.$$

- ▶ Marčenko–Pastur law properties ( $m_\mu(z) = (1 - c - z - czm_\mu(z))^{-1}$ ):

- ▶  $\lambda \mapsto \lambda m_\mu(\lambda) = \int \frac{\lambda}{t-\lambda} \mu(dt)$  maps  $((1 + \sqrt{c})^2, \infty)$  onto  $(-\frac{1+\sqrt{c}}{\sqrt{c}}, 0^-)$
- ▶ Solution only when  $\omega_m > \sqrt{c}$ :

$$\lambda = 1 + \omega_m + c \frac{1 + \omega_m}{\omega_m}.$$



## 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, *finite fourth order moment entries*
- ▶  $C_p = I_p + P$ ,  $P = \sum_{i=1}^K \omega_i u_i u_i^T$ ,  $\omega_1 > \dots > \omega_M > 0$ .

## 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, *finite fourth order moment entries*
- ▶  $C_p = I_p + P$ ,  $P = \sum_{i=1}^K \omega_i u_i u_i^\top$ ,  $\omega_1 > \dots > \omega_M > 0$ .

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

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

In particular,

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

## 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, *finite fourth order moment entries*
- ▶  $C_p = I_p + P$ ,  $P = \sum_{i=1}^K \omega_i u_i u_i^\top$ ,  $\omega_1 > \dots > \omega_M > 0$ .

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

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

In particular,

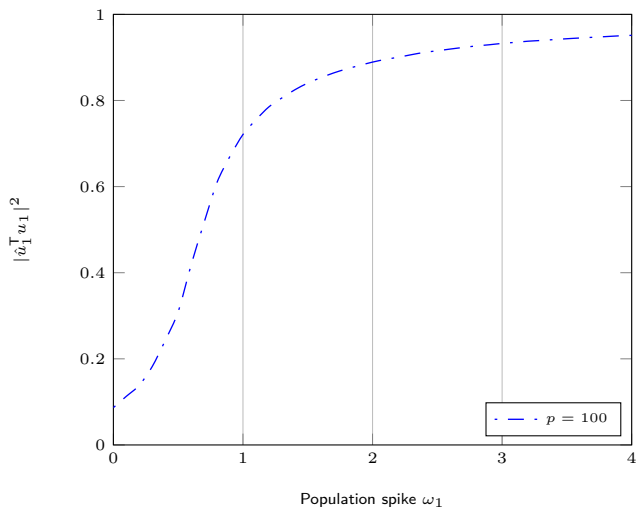
$$|\hat{u}_i^\top u_i|^2 \xrightarrow{\text{a.s.}} \frac{1 - c\omega_i^{-2}}{1 + c\omega_i^{-1}} \cdot \mathbf{1}_{\omega_i > \sqrt{c}}.$$

**Proof:** Based on Cauchy integral + similar ingredients as eigenvalue proof

$$a^\top \hat{u}_i \hat{u}_i^\top b = \frac{1}{2\pi i} \oint_{\mathcal{C}_i} a^\top \left( \frac{1}{n} Y_p Y_p^\top - z I_p \right)^{-1} b dz$$

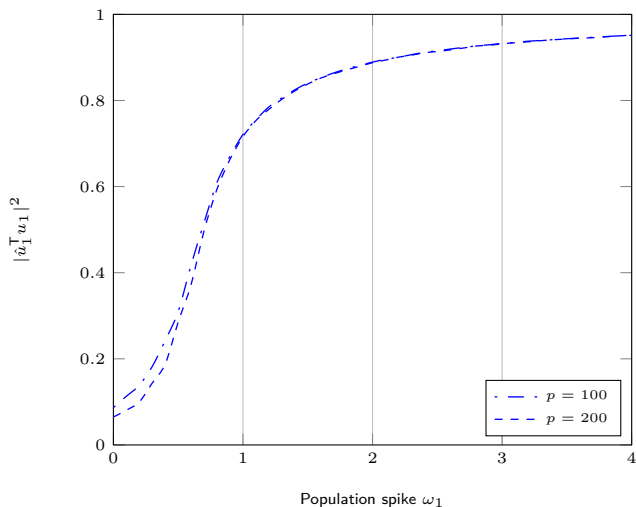
for  $\mathcal{C}_m$  contour circling around  $\lambda_i$  only.

# Spiked Models



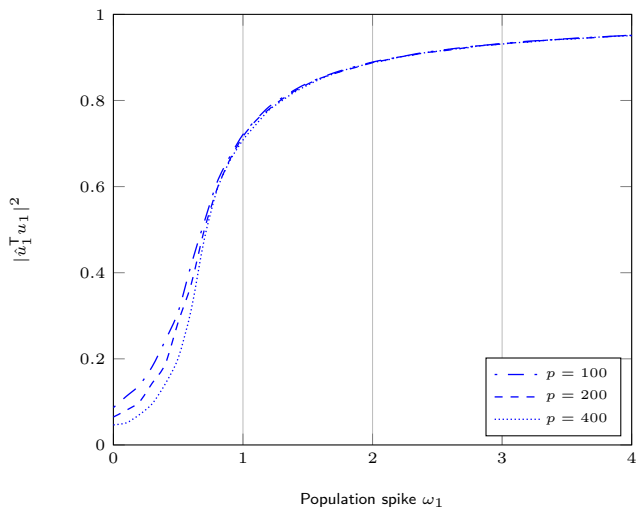
**Figure:** Simulated versus limiting  $|\hat{w}_1^\top u_1|^2$  for  $Y_p = C_p^{\frac{1}{2}} X_p$ ,  $C_p = I_p + \omega_1 u_1 u_1^\top$ ,  $p/n = 1/3$ , varying  $\omega_1$ .

# Spiked Models



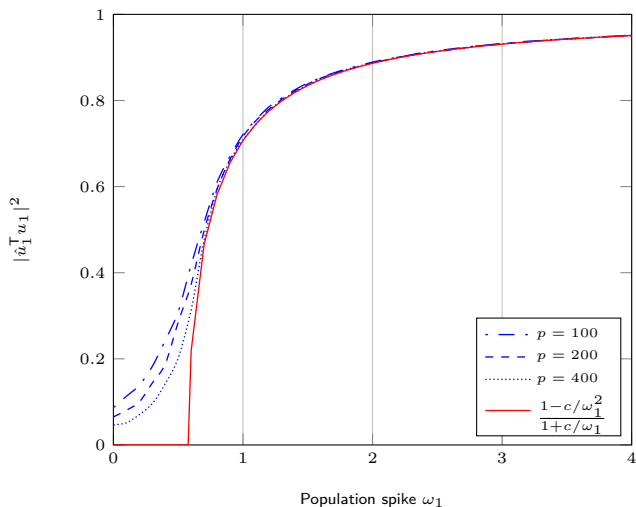
**Figure:** Simulated versus limiting  $|\hat{u}_1^T u_1|^2$  for  $Y_p = C_p^{\frac{1}{2}} X_p$ ,  $C_p = I_p + \omega_1 u_1 u_1^T$ ,  $p/n = 1/3$ , varying  $\omega_1$ .

# Spiked Models



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

## Spiked Models



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



### Theorem (Fluctuations of Eigenvalues [Baik, BenArous, P\'ech\'e'05])

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

- ▶  $X_p$  with i.i.d. *real or complex Gaussian* zero mean, unit variance entries,
- ▶  $C_p = I_p + P$ ,  $P = \sum_{i=1}^K \omega_i u_i u_i^T$ ,  $\omega_1 > \dots > \omega_K > 0$  ( $K \geq 0$ ).

## Tracy–Widom Theorem

### Theorem (Fluctuations of Eigenvalues [Baik, BenArous, P\'ech\'e'05])

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

- ▶  $X_p$  with i.i.d. *real or complex Gaussian* zero mean, unit variance entries,
- ▶  $C_p = I_p + P$ ,  $P = \sum_{i=1}^K \omega_i u_i u_i^T$ ,  $\omega_1 > \dots > \omega_K > 0$  ( $K \geq 0$ ).

Then, as  $p, n \rightarrow \infty$ ,  $p/n \rightarrow c < 1$ ,

- ▶ If  $\omega_1 < \sqrt{c}$  (or  $K = 0$ ),

$$p^{\frac{2}{3}} \frac{\lambda_1 - (1 + \sqrt{c})^2}{(1 + \sqrt{c})^{\frac{4}{3}} c^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} T, \text{ (real or complex Tracy–Widom law)}$$

## Tracy–Widom Theorem

### Theorem (Fluctuations of Eigenvalues [Baik, BenArous, Pécché'05])

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

- ▶  $X_p$  with i.i.d. *real or complex Gaussian* zero mean, unit variance entries,
- ▶  $C_p = I_p + P$ ,  $P = \sum_{i=1}^K \omega_i u_i u_i^T$ ,  $\omega_1 > \dots > \omega_K > 0$  ( $K \geq 0$ ).

Then, as  $p, n \rightarrow \infty$ ,  $p/n \rightarrow c < 1$ ,

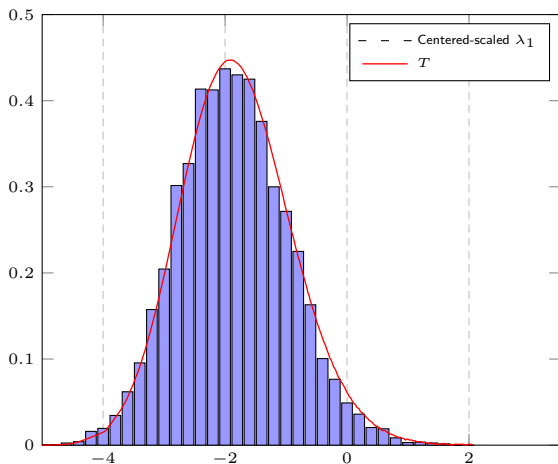
- ▶ If  $\omega_1 < \sqrt{c}$  (or  $K = 0$ ),

$$p^{\frac{2}{3}} \frac{\lambda_1 - (1 + \sqrt{c})^2}{(1 + \sqrt{c})^{\frac{4}{3}} c^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} T, \text{ (real or complex Tracy–Widom law)}$$

- ▶ If  $\omega_1 > \sqrt{c}$ ,

$$\left( \frac{(1 + \omega_1)^2}{c} - \frac{(1 + \omega_1)^2}{\omega_1^2} \right)^{\frac{1}{2}} p^{\frac{1}{2}} \left[ \lambda_1 - \left( 1 + \omega_1 + c \frac{1 + \omega_1}{\omega_1} \right) \right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

## Tracy–Widom Theorem



**Figure:** Distribution of  $p^{\frac{2}{3}} c^{-\frac{1}{2}} (1 + \sqrt{c})^{-\frac{4}{3}} \left[ \lambda_1 \left( \frac{1}{n} X_p X_p^T \right) - (1 + \sqrt{c})^2 \right]$  versus real Tracy–Widom ( $T$ ),  $p = 500$ ,  $n = 1500$ .

Similar results for multiple matrix models:

- ▶  $Y_p = \frac{1}{n}XX^T + P$ ,  $P$  deterministic and low rank
- ▶  $Y_p = \frac{1}{n}X^T(I + P)X$
- ▶  $Y_p = \frac{1}{n}(X + P)^T(X + P)$
- ▶  $Y_p = \frac{1}{n}TX^T(I + P)XT$
- ▶ etc.

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

**Other Common Random Matrix Models**

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

# The Semi-circle law

## Theorem

Let  $X_n \in \mathbb{R}^{n \times n}$  Hermitian with e.s.d.  $\mu_n$  such that  $\frac{1}{\sqrt{n}}[X_n]_{i>j}$  are i.i.d. with zero mean and unit variance. Then, as  $n \rightarrow \infty$ ,

$$\mu_n \xrightarrow{\text{a.s.}} \mu$$

with  $\mu(dt) = \frac{1}{2\pi} \sqrt{(4-t^2)^+} dt$ . In particular,  $m_\mu$  satisfies

$$m_\mu(z) = \frac{1}{-z - m_\mu(z)}.$$

# The Semi-circle law

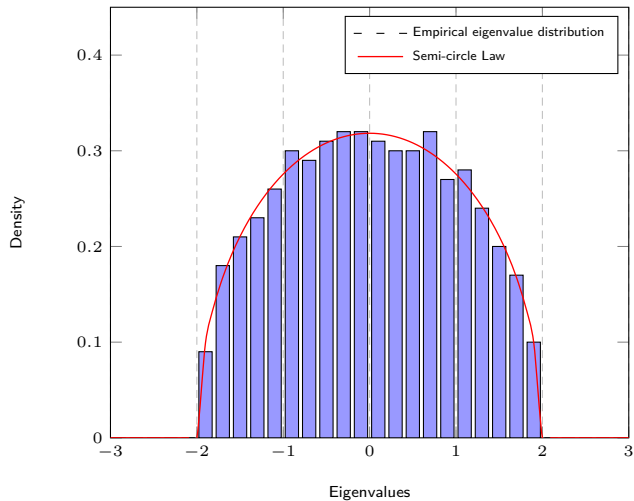


Figure: Histogram of the eigenvalues of Wigner matrices and the semi-circle law, for  $n = 500$



## Theorem

Let  $X_n \in \mathbb{C}^{n \times n}$  with e.s.d.  $\mu_n$  be such that  $\frac{1}{\sqrt{n}}[X_n]_{ij}$  are i.i.d. entries with zero mean and unit variance. Then, as  $n \rightarrow \infty$ ,

$$\mu_n \xrightarrow{\text{a.s.}} \mu$$

with  $\mu$  a complex-supported measure with  $\mu(dz) = \frac{1}{2\pi} \delta_{|z| \leq 1} dz$ .

# The Circular law

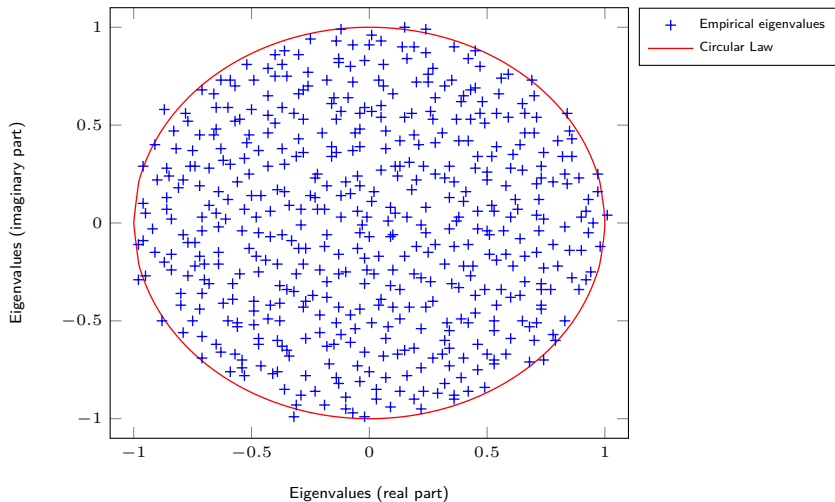







Figure: Eigenvalues of  $X_n$  with i.i.d. standard Gaussian entries, for  $n = 500$ .

### From most accessible to least:

-  Couillet, R., & Debbah, M. (2011). Random matrix methods for wireless communications. Cambridge University Press.
-  Tao, T. (2012). Topics in random matrix theory (Vol. 132). Providence, RI: American Mathematical Society.
-  Bai, Z., & Silverstein, J. W. (2010). Spectral analysis of large dimensional random matrices (Vol. 20). New York: Springer.
-  Pastur, L. A., Shcherbina, M., & Shcherbina, M. (2011). Eigenvalue distribution of large random matrices (Vol. 171). Providence, RI: American Mathematical Society.
-  Anderson, G. W., Guionnet, A., & Zeitouni, O. (2010). An introduction to random matrices (Vol. 118). Cambridge university press.

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

**Applications**

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

**Large range of applications:**

### Large range of applications:

- ▶ **Wireless communications:** capacity of large communication channels  $H \in \mathbb{C}^{p \times n}$ , optimal precoding in mu-MIMO, power allocation in large networks, sensing in cognitive radios, etc.

## From classical applications...

### Large range of applications:

- ▶ **Wireless communications:** capacity of large communication channels  $H \in \mathbb{C}^{p \times n}$ , optimal precoding in mu-MIMO, power allocation in large networks, sensing in cognitive radios, etc.
- ▶ **Array processing:** improved MUSIC methods for large arrays ( $p \sim n$ ), optimal beamforming (MVDR), detection filters (ANMF), etc.

### Large range of applications:

- ▶ **Wireless communications:** capacity of large communication channels  $H \in \mathbb{C}^{p \times n}$ , optimal precoding in mu-MIMO, power allocation in large networks, sensing in cognitive radios, etc.
- ▶ **Array processing:** improved MUSIC methods for large arrays ( $p \sim n$ ), optimal beamforming (MVDR), detection filters (ANMF), etc.
- ▶ **Statistical finance:** portfolio optimization (Markowitz, GMVP) for large portfolios and short time windows.



### Large range of applications:

- ▶ **Wireless communications:** capacity of large communication channels  $H \in \mathbb{C}^{p \times n}$ , optimal precoding in mu-MIMO, power allocation in large networks, sensing in cognitive radios, etc.
- ▶ **Array processing:** improved MUSIC methods for large arrays ( $p \sim n$ ), optimal beamforming (MVDR), detection filters (ANMF), etc.
- ▶ **Statistical finance:** portfolio optimization (Markowitz, GMVP) for large portfolios and short time windows.
- ▶ **Brain signal processing:** EEG covariance estimation on short windows.

### Large range of applications:

- ▶ **Wireless communications:** capacity of large communication channels  $H \in \mathbb{C}^{p \times n}$ , optimal precoding in mu-MIMO, power allocation in large networks, sensing in cognitive radios, etc.
- ▶ **Array processing:** improved MUSIC methods for large arrays ( $p \sim n$ ), optimal beamforming (MVDR), detection filters (ANMF), etc.
- ▶ **Statistical finance:** portfolio optimization (Markowitz, GMVP) for large portfolios and short time windows.
- ▶ **Brain signal processing:** EEG covariance estimation on short windows.

**Any application where  $p \sim n$  “rather large”**

(convergence speed in up to  $O(n)$  and not  $O(\sqrt{n})$  as usual!)

## From classical applications...

### Large range of applications:

- ▶ **Wireless communications:** capacity of large communication channels  $H \in \mathbb{C}^{p \times n}$ , optimal precoding in mu-MIMO, power allocation in large networks, sensing in cognitive radios, etc.
- ▶ **Array processing:** improved MUSIC methods for large arrays ( $p \sim n$ ), optimal beamforming (MVDR), detection filters (ANMF), etc.
- ▶ **Statistical finance:** portfolio optimization (Markowitz, GMVP) for large portfolios and short time windows.
- ▶ **Brain signal processing:** EEG covariance estimation on short windows.

**Any application where  $p \sim n$  “rather large”**

(convergence speed in up to  $O(n)$  and not  $O(\sqrt{n})$  as usual!)

**BUT** mostly linear settings...

**Specificities in statistical and machine learning:**

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.

**Specificities in statistical and machine learning:**

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**Specificities in statistical and machine learning:**

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**CENTRAL ISSUE:** Given that basic sample covariance matrices are not consistent for large  $n, p$ , what happens to machine learning methods?

**Specificities in statistical and machine learning:**

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**CENTRAL ISSUE:** Given that basic sample covariance matrices are not consistent for large  $n, p$ , what happens to machine learning methods?

- ▶ we will see that **small-dimensional intuitions dramatically fail**

**Specificities in statistical and machine learning:**

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**CENTRAL ISSUE:** Given that basic sample covariance matrices are not consistent for large  $n, p$ , what happens to machine learning methods?

- ▶ we will see that **small-dimensional intuitions dramatically fail**
- ▶ some **classical and widely-used algorithms become ineffective**



**Specificities in statistical and machine learning:**

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**CENTRAL ISSUE:** Given that basic sample covariance matrices are not consistent for large  $n, p$ , what happens to machine learning methods?

- ▶ we will see that **small-dimensional intuitions dramatically fail**
- ▶ some **classical and widely-used algorithms become ineffective**
- ▶ **BUT** random matrix theory provides a renewed understanding.

**Specificities in statistical and machine learning:**

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**CENTRAL ISSUE:** Given that basic sample covariance matrices are not consistent for large  $n, p$ , what happens to machine learning methods?

- ▶ we will see that **small-dimensional intuitions dramatically fail**
- ▶ some **classical and widely-used algorithms become ineffective**
- ▶ **BUT** random matrix theory provides a renewed understanding.

**TUTORIAL:** first answers to

## ... to machine learning!

### Specificities in statistical and machine learning:

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**CENTRAL ISSUE:** Given that basic sample covariance matrices are not consistent for large  $n, p$ , what happens to machine learning methods?

- ▶ we will see that **small-dimensional intuitions dramatically fail**
- ▶ some **classical and widely-used algorithms become ineffective**
- ▶ **BUT** random matrix theory provides a renewed understanding.

**TUTORIAL:** first answers to **understand**,

## ... to machine learning!

### Specificities in statistical and machine learning:

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**CENTRAL ISSUE:** Given that basic sample covariance matrices are not consistent for large  $n, p$ , what happens to machine learning methods?

- ▶ we will see that **small-dimensional intuitions dramatically fail**
- ▶ some **classical and widely-used algorithms become ineffective**
- ▶ **BUT** random matrix theory provides a renewed understanding.

**TUTORIAL:** first answers to **understand, improve, and**

## ... to machine learning!

### Specificities in statistical and machine learning:

- ▶ **Matrix of non-linear entries:** kernel matrices  $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ , activation functions in neural nets  $x_{l+1} = \sigma(Wx_l)$ , non-linear features, etc.
- ▶ **Often non-explicit solutions:** robust statistics (fixed-point matrices), SVM margin constraints, logistic regression, etc.

**CENTRAL ISSUE:** Given that basic sample covariance matrices are not consistent for large  $n, p$ , what happens to machine learning methods?

- ▶ we will see that **small-dimensional intuitions dramatically fail**
- ▶ some **classical and widely-used algorithms become ineffective**
- ▶ **BUT** random matrix theory provides a renewed understanding.

**TUTORIAL:** first answers to **understand, improve, and change paradigm**.

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Basics of Random Matrix Theory (**Romain COUILLET**)

- Motivation: Large Sample Covariance Matrices

- The Stieltjes Transform Method

- Spiked Models

- Other Common Random Matrix Models

- Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

- Motivation: EEG-based clustering*

- Covariance Distance Inference

- Revisiting Motivation*

- Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

- Support Vector Machines

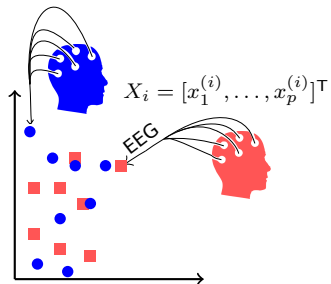
- Semi-Supervised Learning

- From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Motivation example: EEG-based clustering

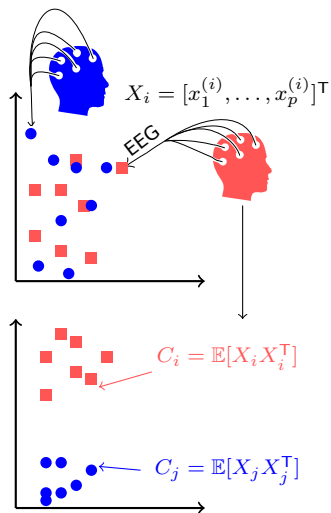
- ▶ Hard classification on raw data  $X_i$ :  
Need Features





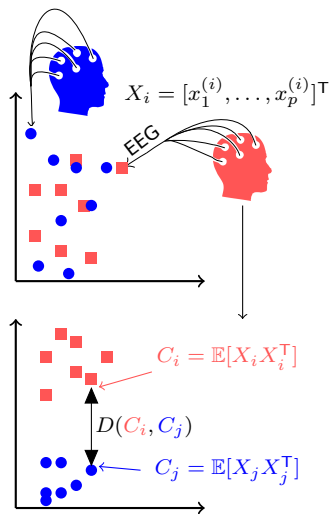
## Motivation example: EEG-based clustering

- ▶ Hard classification on raw data  $X_i$ :  
Need Features
- ▶ Relevant Feature: Covariance  $C_i$



## Motivation example: EEG-based clustering

- ▶ Hard classification on raw data  $X_i$ :  
Need Features
- ▶ Relevant Feature: Covariance  $C_i$
- ▶ Distance between features:  $D(C_i, C_j)$



## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

**Covariance Distance Inference**

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Observations:

- ▶ two data vector classes  $x_i^{(1)} \in \mathcal{C}_1$  and  $x_i^{(2)} \in \mathcal{C}_2$

### Observations:

- ▶ two data vector classes  $x_i^{(1)} \in \mathcal{C}_1$  and  $x_i^{(2)} \in \mathcal{C}_2$
- ▶  $X_a = [x_1^{(a)}, \dots, x_{n_a}^{(a)}]$ ,  $x_i^{(a)} \in \mathbb{R}^p$  with  $E[x_i^{(a)}] = 0$ ,  $E[x_i^{(a)} x_i^{(a)\top}] = C_a$ .

## Observations:

- ▶ two data vector classes  $x_i^{(1)} \in C_1$  and  $x_i^{(2)} \in C_2$
- ▶  $X_a = [x_1^{(a)}, \dots, x_{n_a}^{(a)}]$ ,  $x_i^{(a)} \in \mathbb{R}^p$  with  $E[x_i^{(a)}] = 0$ ,  $E[x_i^{(a)} x_i^{(a)\top}] = C_a$ .

## Objective:

- ▶ From the data  $x_i^{(a)}$ , estimate some distance function

$$D \equiv D(C_1, C_2).$$

## Observations:

- ▶ two data vector classes  $x_i^{(1)} \in \mathcal{C}_1$  and  $x_i^{(2)} \in \mathcal{C}_2$
- ▶  $X_a = [x_1^{(a)}, \dots, x_{n_a}^{(a)}]$ ,  $x_i^{(a)} \in \mathbb{R}^p$  with  $E[x_i^{(a)}] = 0$ ,  $E[x_i^{(a)} x_i^{(a)\top}] = C_a$ .

## Objective:

- ▶ From the data  $x_i^{(a)}$ , estimate some distance function

$$D \equiv D(\mathcal{C}_1, \mathcal{C}_2).$$

- ▶ Classical approach:

$$\hat{D} \equiv D(\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2), \quad \text{with } \hat{C}_a = \frac{1}{n_a} \sum_{i=1}^{n_a} x_i^{(a)} x_i^{(a)\top} = \frac{1}{n_a} X_a X_a^\top.$$

### Observations:

- ▶ two data vector classes  $x_i^{(1)} \in \mathcal{C}_1$  and  $x_i^{(2)} \in \mathcal{C}_2$
- ▶  $X_a = [x_1^{(a)}, \dots, x_{n_a}^{(a)}]$ ,  $x_i^{(a)} \in \mathbb{R}^p$  with  $E[x_i^{(a)}] = 0$ ,  $E[x_i^{(a)} x_i^{(a)\top}] = C_a$ .

### Objective:

- ▶ From the data  $x_i^{(a)}$ , estimate some distance function

$$D \equiv D(C_1, C_2).$$

- ▶ Classical approach:

$$\hat{D} \equiv D(\hat{C}_1, \hat{C}_2), \quad \text{with } \hat{C}_a = \frac{1}{n_a} \sum_{i=1}^{n_a} x_i^{(a)} x_i^{(a)\top} = \frac{1}{n_a} X_a X_a^\top.$$

→ Often justified by Law of Large Numbers:  $\hat{D} \xrightarrow{\text{a.s.}} D$  as  $n \rightarrow \infty$ .



## In practice though...

### Example:

- ▶ The Fisher distance

$$D(C_1, C_2) = \frac{1}{p} \left\| \log^2(C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}) \right\|_F^2$$

## In practice though...

### Example:

- ▶ The Fisher distance

$$D(C_1, C_2) = \frac{1}{p} \left\| \log^2(C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}) \right\|_F^2 = \frac{1}{p} \sum_{i=1}^p \log^2(\lambda_i(C_1^{-1} C_2))$$

## In practice though...

### Example:

- ▶ The Fisher distance

$$D(C_1, C_2) = \frac{1}{p} \left\| \log^2(C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}) \right\|_F^2 = \frac{1}{p} \sum_{i=1}^p \log^2(\lambda_i(C_1^{-1} C_2)) = \int \log^2(t) \nu_p(dt)$$

with  $\nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1} C_2)}$ .

## In practice though...

### Example:

- ▶ The Fisher distance

$$D(C_1, C_2) = \frac{1}{p} \left\| \log^2(C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}) \right\|_F^2 = \frac{1}{p} \sum_{i=1}^p \log^2(\lambda_i(C_1^{-1} C_2)) = \int \log^2(t) \nu_p(dt)$$

$$\text{with } \nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1} C_2)}.$$

- ▶ for  $n_1 = 1024$ ,  $n_2 = 2048$ , **different**  $p$  (here  $[C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}]_{ij} = .3^{|i-j|}$ ):

## In practice though...

### Example:

- ▶ The Fisher distance

$$D(C_1, C_2) = \frac{1}{p} \left\| \log^2(C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}) \right\|_F^2 = \frac{1}{p} \sum_{i=1}^p \log^2(\lambda_i(C_1^{-1} C_2)) = \int \log^2(t) \nu_p(dt)$$

$$\text{with } \nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1} C_2)}.$$

- ▶ for  $n_1 = 1024$ ,  $n_2 = 2048$ , **different**  $p$  (here  $[C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}]_{ij} = .3^{|i-j|}$ ):

$p$	Fisher distance	Classical estimator
2	0.0980	<b>0.1002</b>
4	0.1456	<b>0.1520</b>
8	0.1694	0.1820
16	0.1812	0.2081
32	0.1872	0.2363
64	0.1901	<b>0.2892</b>
128	0.1916	<b>0.3955</b>
256	0.1924	<b>0.6338</b>
512	0.1927	<b><u>1.2715</u></b>

(error < 5%) (error > 50%) (error > 100%) (error > 500%)

## In practice though...

### Example:

- ▶ The Fisher distance

$$D(C_1, C_2) = \frac{1}{p} \left\| \log^2(C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}) \right\|_F^2 = \frac{1}{p} \sum_{i=1}^p \log^2(\lambda_i(C_1^{-1} C_2)) = \int \log^2(t) \nu_p(dt)$$

$$\text{with } \nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1} C_2)}.$$

- ▶ for  $n_1 = 1024$ ,  $n_2 = 2048$ , **different**  $p$  (here  $[C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}]_{ij} = .3^{|i-j|}$ ):

$p$	Fisher distance	Classical estimator	RMT estimator
2	0.0980	0.1002	0.0973
4	0.1456	0.1520	0.1461
8	0.1694	0.1820	0.1703
16	0.1812	0.2081	0.1845
32	0.1872	0.2363	0.1886
64	0.1901	<b>0.2892</b>	0.1920
128	0.1916	<b>0.3955</b>	0.1934
256	0.1924	<b>0.6338</b>	0.1942
512	0.1927	<b>1.2715</b>	0.1953

(error < 5%) (error > 50%) (error > 100%) (error > 500%)

## Explanation for failure

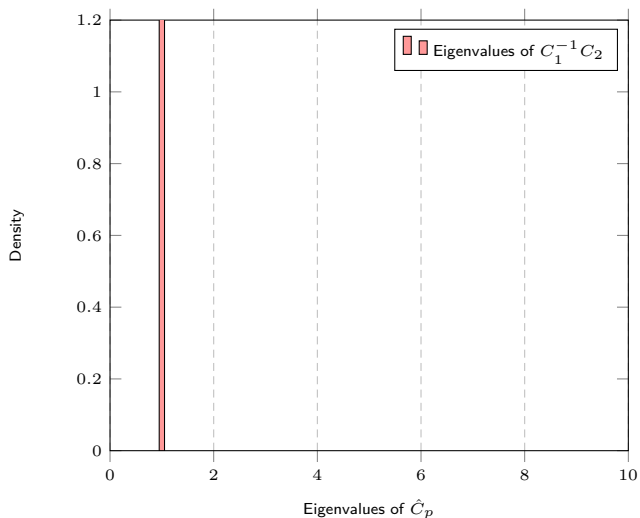


Figure: Population and Sample Eigenvalues for  $n_1 = 1024$ ,  $n_2 = 2048$ , varying  $p$ ,  $C_1 = C_2$ .

## Explanation for failure

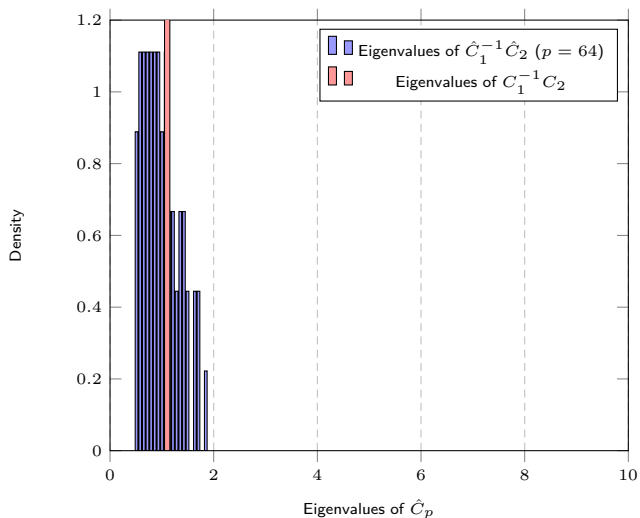


Figure: Population and Sample Eigenvalues for  $n_1 = 1024$ ,  $n_2 = 2048$ , varying  $p$ ,  $C_1 = C_2$ .



## Explanation for failure

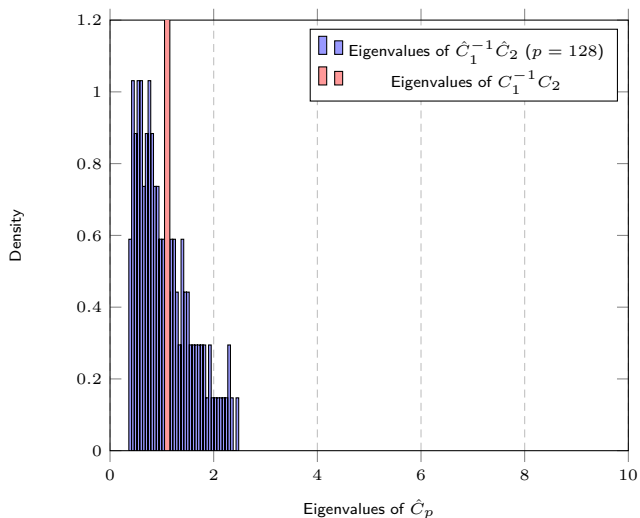


Figure: Population and Sample Eigenvalues for  $n_1 = 1024$ ,  $n_2 = 2048$ , varying  $p$ ,  $C_1 = C_2$ .

## Explanation for failure

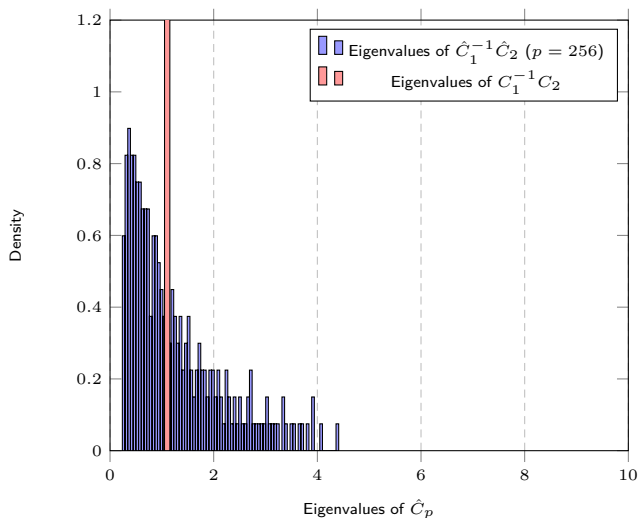


Figure: Population and Sample Eigenvalues for  $n_1 = 1024$ ,  $n_2 = 2048$ , varying  $p$ ,  $C_1 = C_2$ .

## Explanation for failure

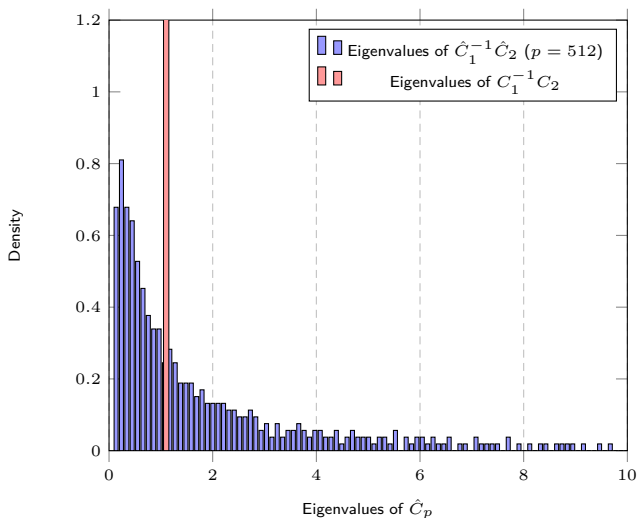


Figure: Population and Sample Eigenvalues for  $n_1 = 1024$ ,  $n_2 = 2048$ , varying  $p$ ,  $C_1 = C_2$ .

## Assumptions

- ▶ **[Spatial independence]**  $x_i^{(a)} = C_a^{\frac{1}{2}} \tilde{x}_i^{(a)}$ ,  $\tilde{x}_i^{(a)} \in \mathbb{R}^p$  with i.i.d. zero mean, unit variance, finite  $4 + \varepsilon$  order moment.

## Assumptions

▶ **[Spatial independence]**  $x_i^{(a)} = C_a^{\frac{1}{2}} \tilde{x}_i^{(a)}$ ,  $\tilde{x}_i^{(a)} \in \mathbb{R}^p$  with i.i.d. zero mean, unit variance, finite  $4 + \varepsilon$  order moment.

▶ **[RMT regime]** As  $n_a \rightarrow \infty$ ,

$$\frac{p}{n_a} = c_a \rightarrow c_a^\infty \in (0, 1).$$

▶ **[Studied distances]** for  $f$  a complex-analytic extensible function,

$$D(C_1, C_2) = \int f(t) \nu_p(dt)$$

## Assumptions

- ▶ **[Spatial independence]**  $x_i^{(a)} = C_a^{\frac{1}{2}} \tilde{x}_i^{(a)}$ ,  $\tilde{x}_i^{(a)} \in \mathbb{R}^p$  with i.i.d. zero mean, unit variance, finite  $4 + \varepsilon$  order moment.

- ▶ **[RMT regime]** As  $n_a \rightarrow \infty$ ,

$$\frac{p}{n_a} = c_a \rightarrow c_a^\infty \in (0, 1).$$

- ▶ **[Studied distances]** for  $f$  a complex-analytic extensible function,

$$D(C_1, C_2) = \int f(t) \nu_p(dt), \quad \nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1}C_2)}.$$

## Assumptions

- ▶ **[Spatial independence]**  $x_i^{(a)} = C_a^{\frac{1}{2}} \tilde{x}_i^{(a)}$ ,  $\tilde{x}_i^{(a)} \in \mathbb{R}^p$  with i.i.d. zero mean, unit variance, finite  $4 + \varepsilon$  order moment.

- ▶ **[RMT regime]** As  $n_a \rightarrow \infty$ ,

$$\frac{p}{n_a} = c_a \rightarrow c_a^\infty \in (0, 1).$$

- ▶ **[Studied distances]** for  $f$  a complex-analytic extensible function,

$$D(C_1, C_2) = \int f(t) \nu_p(dt), \quad \nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1} C_2)}.$$

## Examples

- ▶ *Fisher geodesic distance:*  $f(t) = \log^2(t)$
- ▶ *Bhattacharyya distance:*  $f(t) = -\frac{1}{4} \log(t) + \frac{1}{2} \log(1+t) - \frac{1}{2} \log(2)$
- ▶ *Kullback-Leibler divergence for Gaussian:*  $f(t) = \frac{1}{2} t - \frac{1}{2} \log(t) - \frac{1}{2}$
- ▶ *Rényi divergence for Gaussian:*  $f(t) = \frac{-1}{2(\alpha-1)} \log(\alpha + (1-\alpha)t) + \frac{1}{2} \log(t)$

**Notations:**



## RMT-improved estimator

### Notations:

- ▶ Population eigenvalue distribution:  $\nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1}C_2)}$

### Notations:

- ▶ **Population eigenvalue distribution:**  $\nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1}C_2)}$
- ▶ **Sample eigenvalue distribution:**  $\mu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(\hat{C}_1^{-1}\hat{C}_2)}$

### Notations:

- ▶ **Population eigenvalue distribution:**  $\nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1}C_2)}$
- ▶ **Sample eigenvalue distribution:**  $\mu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(\hat{C}_1^{-1}\hat{C}_2)} \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i}$

### Notations:

- ▶ **Population eigenvalue distribution:**  $\nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1}C_2)}$
- ▶ **Sample eigenvalue distribution:**  $\mu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(\hat{C}_1^{-1}\hat{C}_2)} \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i}$
- ▶ Recall **Stieltjes transform**  $m_\theta(z)$ ,  $z \in \mathbb{C} \setminus \text{Supp}(\theta)$ , of measure  $\theta$ :

$$m_\theta(z) = \int \frac{1}{\lambda - z} d\theta(\lambda)$$

## Notations:

- ▶ **Population eigenvalue distribution:**  $\nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1}C_2)}$
- ▶ **Sample eigenvalue distribution:**  $\mu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(\hat{C}_1^{-1}\hat{C}_2)} \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i}$
- ▶ Recall **Stieltjes transform**  $m_\theta(z)$ ,  $z \in \mathbb{C} \setminus \text{Supp}(\theta)$ , of measure  $\theta$ :

$$m_\theta(z) = \int \frac{1}{\lambda - z} d\theta(\lambda)$$

e.g.,  $m_{\mu_p}(z) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i - z}$ .

## Notations:

- ▶ **Population eigenvalue distribution:**  $\nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1}C_2)}$
- ▶ **Sample eigenvalue distribution:**  $\mu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(\hat{C}_1^{-1}\hat{C}_2)} \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i}$
- ▶ Recall **Stieltjes transform**  $m_\theta(z)$ ,  $z \in \mathbb{C} \setminus \text{Supp}(\theta)$ , of measure  $\theta$ :

$$m_\theta(z) = \int \frac{1}{\lambda - z} d\theta(\lambda)$$

$$\text{e.g., } m_{\mu_p}(z) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i - z}.$$

## Theorem (Estimation via contour integral)

For  $z \in \mathbb{C} \setminus \text{Supp}(\mu_p)$ , let

$$\varphi_p(z) \equiv z + c_1 z^2 m_{\mu_p}(z)$$

$$\psi_p(z) \equiv 1 - c_2 - c_2 z m_{\mu_p}(z).$$

## RMT-improved estimator

### Notations:

- ▶ **Population eigenvalue distribution:**  $\nu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(C_1^{-1}C_2)}$
- ▶ **Sample eigenvalue distribution:**  $\mu_p = \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(\hat{C}_1^{-1}\hat{C}_2)} \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i}$
- ▶ Recall **Stieltjes transform**  $m_\theta(z)$ ,  $z \in \mathbb{C} \setminus \text{Supp}(\theta)$ , of measure  $\theta$ :

$$m_\theta(z) = \int \frac{1}{\lambda - z} d\theta(\lambda)$$

$$\text{e.g., } m_{\mu_p}(z) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i - z}.$$

### Theorem (Estimation via contour integral)

For  $z \in \mathbb{C} \setminus \text{Supp}(\mu_p)$ , let

$$\varphi_p(z) \equiv z + c_1 z^2 m_{\mu_p}(z)$$

$$\psi_p(z) \equiv 1 - c_2 - c_2 z m_{\mu_p}(z).$$

Then, for any (positively oriented) contour  $\Gamma \subset \{z \in \mathbb{C}, \Re[z] > 0\}$  surrounding  $\text{Supp}(\mu_p)$ .

$$\int f d\nu_p - \frac{1}{2\pi i} \oint_{\Gamma} f \left( \frac{\varphi_p(z)}{\psi_p(z)} \right) \left( \frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)} \right) \frac{\psi_p(z)}{c_2} dz \xrightarrow{\text{a.s.}} 0.$$

## Idea of proof

From [Bai-Silverstein'95]<sup>1</sup>, limiting spectra of  $C$  and  $\hat{C}$  related through Stieljes transform.

---

<sup>1</sup>SIL95.



## Idea of proof

From [Bai-Silverstein'95]<sup>1</sup>, limiting spectra of  $C$  and  $\hat{C}$  related through Stieljes transform.

Besides, by Cauchy's integral,

$$\int f(t)\nu_p(dt) = \int \left[ \frac{-1}{2\pi i} \oint_{\Gamma} \frac{f(z)}{t-z} dz \right] \nu_p(dt)$$

---

<sup>1</sup>SIL95.

## Idea of proof

From [Bai-Silverstein'95]<sup>1</sup>, limiting spectra of  $C$  and  $\hat{C}$  related through Stieljes transform.

Besides, by Cauchy's integral,

$$\int f(t)\nu_p(dt) = \int \left[ \frac{-1}{2\pi i} \oint_{\Gamma} \frac{f(z)}{t-z} dz \right] \nu_p(dt) = \frac{-1}{2\pi i} \oint_{\Gamma} f(z) \underbrace{\left[ \int \frac{1}{t-z} \nu_p(dt) \right]}_{=m_{\nu_p}(z)} dz.$$

---

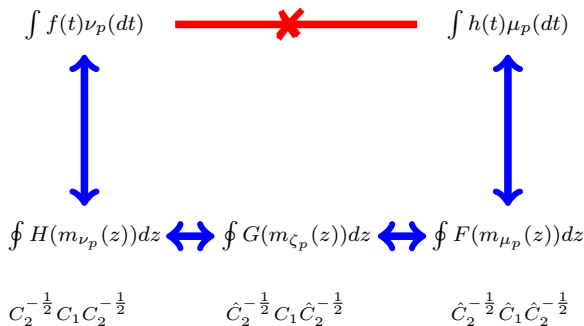
<sup>1</sup>SIL95.

## Idea of proof

From [Bai-Silverstein'95]<sup>1</sup>, limiting spectra of  $C$  and  $\hat{C}$  related through Stieljes transform.

Besides, by Cauchy's integral,

$$\int f(t)\nu_p(dt) = \int \left[ \frac{-1}{2\pi i} \oint_{\Gamma} \frac{f(z)}{t-z} dz \right] \nu_p(dt) = \frac{-1}{2\pi i} \oint_{\Gamma} f(z) \underbrace{\left[ \int \frac{1}{t-z} \nu_p(dt) \right]}_{=m_{\nu_p}(z)} dz.$$



<sup>1</sup>SIL95.

## Evaluation of the complex integrals

**Object of interest:** Evaluate in closed-form

$$\frac{1}{2\pi i} \oint_{\Gamma} f \left( \frac{\varphi_p(z)}{\psi_p(z)} \right) \left( \frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)} \right) \frac{\psi_p(z)}{c_2} dz.$$

# Evaluation of the complex integrals

**Object of interest:** Evaluate in closed-form

$$\frac{1}{2\pi i} \oint_{\Gamma} f \left( \frac{\varphi_p(z)}{\psi_p(z)} \right) \left( \frac{\varphi'_p(z)}{\varphi_p(z)} - \frac{\psi'_p(z)}{\psi_p(z)} \right) \frac{\psi_p(z)}{c_2} dz.$$

**Reminder:** functions of interest

- ▶ *Fisher geodesic distance:*  $f(t) = \log^2(t)$
- ▶ *Bhattacharyya distance:*  $f(t) = -\frac{1}{4} \log(t) + \frac{1}{2} \log(1+t) - \frac{1}{2} \log(2)$
- ▶ *Kullback-Leibler divergence for Gaussian:*  $f(t) = \frac{1}{2} t - \frac{1}{2} \log(t) - \frac{1}{2}$
- ▶ *Rényi divergence for Gaussian:*  $f(t) = \frac{-1}{2(\alpha-1)} \log(\alpha + (1-\alpha)t) + \frac{1}{2} \log(t)$

# Evaluation of the complex integrals

**Object of interest:** Evaluate in closed-form

$$\frac{1}{2\pi i} \oint_{\Gamma} f \left( \frac{\varphi_p(z)}{\psi_p(z)} \right) \left( \frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)} \right) \frac{\psi_p(z)}{c_2} dz.$$

**Reminder:** functions of interest

- ▶ *Fisher geodesic distance:*  $f(t) = \log^2(t)$
- ▶ *Bhattacharyya distance:*  $f(t) = -\frac{1}{4} \log(t) + \frac{1}{2} \log(1+t) - \frac{1}{2} \log(2)$
- ▶ *Kullback-Leibler divergence for Gaussian:*  $f(t) = \frac{1}{2} t - \frac{1}{2} \log(t) - \frac{1}{2}$
- ▶ *Rényi divergence for Gaussian:*  $f(t) = \frac{-1}{2(\alpha-1)} \log(\alpha + (1-\alpha)t) + \frac{1}{2} \log(t)$

**Cases of interest:**

- ▶ Entire functions (e.g.,  $f(t) = t$ ): **residue calculus**

# Evaluation of the complex integrals

**Object of interest:** Evaluate in closed-form

$$\frac{1}{2\pi i} \oint_{\Gamma} f \left( \frac{\varphi_p(z)}{\psi_p(z)} \right) \left( \frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)} \right) \frac{\psi_p(z)}{c_2} dz.$$

**Reminder:** functions of interest

- ▶ *Fisher geodesic distance:*  $f(t) = \log^2(t)$
- ▶ *Bhattacharyya distance:*  $f(t) = -\frac{1}{4} \log(t) + \frac{1}{2} \log(1+t) - \frac{1}{2} \log(2)$
- ▶ *Kullback-Leibler divergence for Gaussian:*  $f(t) = \frac{1}{2} t - \frac{1}{2} \log(t) - \frac{1}{2}$
- ▶ *Rényi divergence for Gaussian:*  $f(t) = \frac{-1}{2(\alpha-1)} \log(\alpha + (1-\alpha)t) + \frac{1}{2} \log(t)$

**Cases of interest:**

- ▶ Entire functions (e.g.,  $f(t) = t$ ): **residue calculus**
- ▶ Functions with **branch cuts:**  $f(t) = \log(t)$ ,  $f(t) = \log(1+st)$ ,  $f(t) = \log^2(t)$ , etc.  
→ **Much more technical!**

**The case**  $f(t) = \log^k(t)$

- ▶ Much less trivial due to **branch cuts of  $\log(z)$** !!

$$\log(z) \equiv \log(|z|) + i \arg(z), \quad \arg(z) \in (-\pi, \pi].$$



**The case**  $f(t) = \log^k(t)$

- ▶ Much less trivial due to **branch cuts of  $\log(z)$ !!**

$$\log(z) \equiv \log(|z|) + i \arg(z), \quad \arg(z) \in (-\pi, \pi].$$

- ▶ Singularities arise when  $\log(\varphi_p(z)/\psi_p(z))$  **discontinuous**.

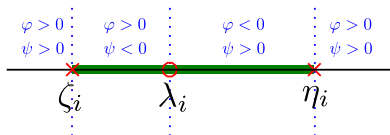
## Sketch of Proof

The case  $f(t) = \log^k(t)$

- ▶ Much less trivial due to **branch cuts of  $\log(z)$ !!**

$$\log(z) \equiv \log(|z|) + i \arg(z), \quad \arg(z) \in (-\pi, \pi].$$

- ▶ Singularities arise when  $\log(\varphi_p(z)/\psi_p(z))$  **discontinuous**.
- ▶ The situation in image...



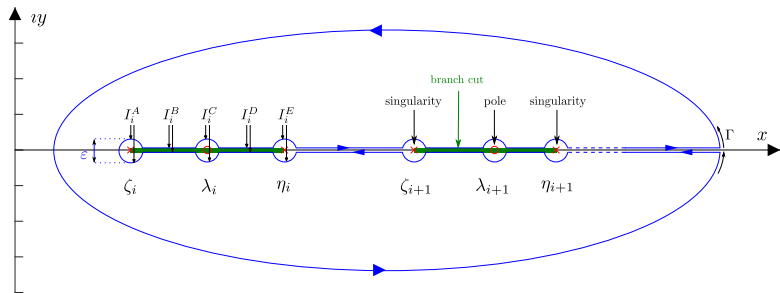
with

- ▶  $\zeta_i$  zeros of  $\psi_p$
- ▶  $\eta_i$  zeros of  $\varphi_p$ .

## Sketch of proof

The case  $f(t) = \log^k(t)$  (continued)

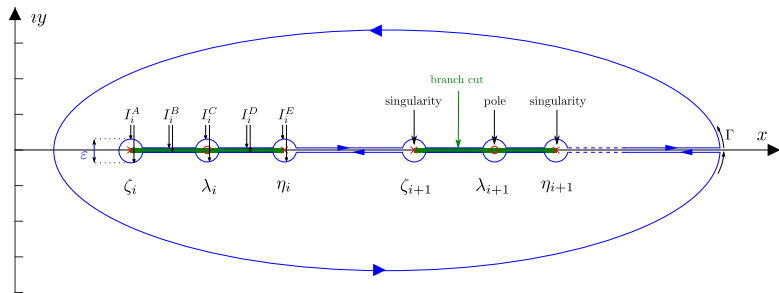
- Integration method: avoid branch cuts:



## Sketch of proof

The case  $f(t) = \log^k(t)$  (continued)

- Integration method: avoid branch cuts:



- Detailed method:

- careful control of integrals on circles  $I_i^A$ ,  $I_i^C$ ,  $I_i^E$  (Jordan's identity does not apply!)
- linear integrals on segments, up to integrability... **easy for  $\log(t)$ , difficult for  $\log^2(t)$ !**

Corollary (Case  $f(t) = t$ )

*Under the same assumptions,*

$$\int t\nu_p(dt) - (1 - c_1) \int t\mu_p(dt) \xrightarrow{\text{a.s.}} 0.$$

## Application to specific functions

Corollary (Case  $f(t) = t$ )

*Under the same assumptions,*

$$\int t\nu_p(dt) - (1 - c_1) \int t\mu_p(dt) \xrightarrow{\text{a.s.}} 0.$$

$$\text{(i.e., } \frac{1}{p} \text{tr } C_1^{-1} C_2 \simeq (1 - \frac{p}{n_1}) \frac{1}{p} \text{tr } \hat{C}_1^{-1} \hat{C}_2 \text{)}$$

## Application to specific functions

Corollary (Case  $f(t) = t$ )

*Under the same assumptions,*

$$\int t\nu_p(dt) - (1 - c_1) \int t\mu_p(dt) \xrightarrow{\text{a.s.}} 0.$$

(i.e.,  $\frac{1}{p} \text{tr} C_1^{-1} C_2 \simeq (1 - \frac{p}{n_1}) \frac{1}{p} \text{tr} \hat{C}_1^{-1} \hat{C}_2$ )

→ Just a scaling factor!

## Application to specific functions

Corollary (Case  $f(t) = t$ )

*Under the same assumptions,*

$$\int t\nu_p(dt) - (1 - c_1) \int t\mu_p(dt) \xrightarrow{\text{a.s.}} 0.$$

$$\text{(i.e., } \frac{1}{p} \text{tr } C_1^{-1} C_2 \simeq (1 - \frac{p}{n_1}) \frac{1}{p} \text{tr } \hat{C}_1^{-1} \hat{C}_2 \text{)}$$

→ Just a scaling factor!

Corollary (Case  $f(t) = \log(t)$ )

*Under the same assumptions,*

$$\int \log(t)\nu_p(dt) - \left[ \int \log(t)\mu_p(dt) - \frac{1 - c_1}{c_1} \log(1 - c_1) + \frac{1 - c_2}{c_2} \log(1 - c_2) \right] \xrightarrow{\text{a.s.}} 0.$$



## Application to specific functions

### Corollary (Case $f(t) = t$ )

Under the same assumptions,

$$\int t\nu_p(dt) - (1 - c_1) \int t\mu_p(dt) \xrightarrow{\text{a.s.}} 0.$$

$$\text{(i.e., } \frac{1}{p} \text{tr } C_1^{-1} C_2 \simeq (1 - \frac{p}{n_1}) \frac{1}{p} \text{tr } \hat{C}_1^{-1} \hat{C}_2)$$

→ Just a scaling factor!

### Corollary (Case $f(t) = \log(t)$ )

Under the same assumptions,

$$\int \log(t)\nu_p(dt) - \left[ \int \log(t)\mu_p(dt) - \frac{1 - c_1}{c_1} \log(1 - c_1) + \frac{1 - c_2}{c_2} \log(1 - c_2) \right] \xrightarrow{\text{a.s.}} 0.$$

$$\text{(i.e., } \frac{1}{p} \log \det(C_1^{-1} C_2) \simeq \frac{1}{p} \log \det(\hat{C}_1^{-1} \hat{C}_2) - \frac{n_1 - p}{n_1} \log(1 - \frac{p}{n_1}) + \frac{n_2 - p}{n_2} \log(1 - \frac{p}{n_2}))$$

## Application to specific functions

Corollary (Case  $f(t) = t$ )

Under the same assumptions,

$$\int t\nu_p(dt) - (1 - c_1) \int t\mu_p(dt) \xrightarrow{\text{a.s.}} 0.$$

$$\text{(i.e., } \frac{1}{p} \text{tr } C_1^{-1} C_2 \simeq (1 - \frac{p}{n_1}) \frac{1}{p} \text{tr } \hat{C}_1^{-1} \hat{C}_2)$$

→ Just a scaling factor!

Corollary (Case  $f(t) = \log(t)$ )

Under the same assumptions,

$$\int \log(t)\nu_p(dt) - \left[ \int \log(t)\mu_p(dt) - \frac{1 - c_1}{c_1} \log(1 - c_1) + \frac{1 - c_2}{c_2} \log(1 - c_2) \right] \xrightarrow{\text{a.s.}} 0.$$

$$\text{(i.e., } \frac{1}{p} \log \det(C_1^{-1} C_2) \simeq \frac{1}{p} \log \det(\hat{C}_1^{-1} \hat{C}_2) - \frac{n_1 - p}{n_1} \log(1 - \frac{p}{n_1}) + \frac{n_2 - p}{n_2} \log(1 - \frac{p}{n_2}))$$

→ Just a bias term!

### Corollary (Case $f(t) = \log(1 + st)$ )

Denoting  $\kappa_0 < 0$  unique negative solution to  $1 + s \frac{\varphi_p(x)}{\psi_p(x)} = 0$ ,

$$\int \log(1 + st) d\nu_p(t) - \left[ \frac{c_1 + c_2 - c_1 c_2}{c_1 c_2} \log \left( \frac{c_1 + c_2 - c_1 c_2}{(1 - c_1)(c_2 - s c_1 \kappa_0)} \right) + \frac{1}{c_2} \log(-s \kappa_0 (1 - c_1)) + \int \log \left( 1 - \frac{t}{\kappa_0} \right) d\mu_p(t) \right] \xrightarrow{\text{a.s.}} 0.$$

Corollary (Case  $f(t) = \log(1 + st)$ )

Denoting  $\kappa_0 < 0$  unique negative solution to  $1 + s \frac{\varphi_p(x)}{\psi_p(x)} = 0$ ,

$$\int \log(1 + st) d\nu_p(t) - \left[ \frac{c_1 + c_2 - c_1 c_2}{c_1 c_2} \log \left( \frac{c_1 + c_2 - c_1 c_2}{(1 - c_1)(c_2 - s c_1 \kappa_0)} \right) + \frac{1}{c_2} \log(-s \kappa_0 (1 - c_1)) + \int \log \left( 1 - \frac{t}{\kappa_0} \right) d\mu_p(t) \right] \xrightarrow{\text{a.s.}} 0.$$

→ Highly non-trivial!

Corollary (Case  $f(t) = \log^2(t)$ )

$$\begin{aligned}
 & \frac{1}{2\pi i} \oint_{\Gamma} \log^2 \left( \frac{\varphi_p(z)}{\psi_p(z)} \right) \left( \frac{\varphi'_p(z)}{\varphi_p(z)} - \frac{\psi'_p(z)}{\psi_p(z)} \right) \frac{\psi_p(z)}{c_2} dz \\
 &= \frac{c_1 + c_2 - c_1 c_2}{c_1 c_2} \left[ \sum_{i=1}^p \left\{ \log^2((1 - c_1)\eta_i) - \log^2((1 - c_1)\lambda_i) \right\} \right. \\
 &+ 2 \sum_{1 \leq i, j \leq p} \left\{ \operatorname{Li}_2 \left( 1 - \frac{\zeta_i}{\lambda_j} \right) - \operatorname{Li}_2 \left( 1 - \frac{\eta_i}{\lambda_j} \right) + \operatorname{Li}_2 \left( 1 - \frac{\eta_i}{\eta_j} \right) - \operatorname{Li}_2 \left( 1 - \frac{\zeta_i}{\eta_j} \right) \right\} \left. \right] \\
 &- \frac{1 - c_2}{c_2} \left[ \log^2(1 - c_2) - \log^2(1 - c_1) + \sum_{i=1}^p \left\{ \log^2(\eta_i) - \log^2(\zeta_i) \right\} \right] \\
 &- \frac{1}{p} \left[ 2 \sum_{1 \leq i, j \leq p} \left\{ \operatorname{Li}_2 \left( 1 - \frac{\zeta_i}{\lambda_j} \right) - \operatorname{Li}_2 \left( 1 - \frac{\eta_i}{\lambda_j} \right) \right\} - \sum_{i=1}^p \log^2((1 - c_1)\lambda_i) \right]
 \end{aligned}$$

## Application to specific functions

Corollary (Case  $f(t) = \log^2(t)$ )

$$\begin{aligned} & \frac{1}{2\pi i} \oint_{\Gamma} \log^2 \left( \frac{\varphi_p(z)}{\psi_p(z)} \right) \left( \frac{\varphi'_p(z)}{\varphi_p(z)} - \frac{\psi'_p(z)}{\psi_p(z)} \right) \frac{\psi_p(z)}{c_2} dz \\ &= \frac{c_1 + c_2 - c_1 c_2}{c_1 c_2} \left[ \sum_{i=1}^p \left\{ \log^2((1 - c_1)\eta_i) - \log^2((1 - c_1)\lambda_i) \right\} \right. \\ &+ 2 \sum_{1 \leq i, j \leq p} \left\{ \operatorname{Li}_2 \left( 1 - \frac{\zeta_i}{\lambda_j} \right) - \operatorname{Li}_2 \left( 1 - \frac{\eta_i}{\lambda_j} \right) + \operatorname{Li}_2 \left( 1 - \frac{\eta_i}{\eta_j} \right) - \operatorname{Li}_2 \left( 1 - \frac{\zeta_i}{\eta_j} \right) \right\} \left. \right] \\ &- \frac{1 - c_2}{c_2} \left[ \log^2(1 - c_2) - \log^2(1 - c_1) + \sum_{i=1}^p \left\{ \log^2(\eta_i) - \log^2(\zeta_i) \right\} \right] \\ &- \frac{1}{p} \left[ 2 \sum_{1 \leq i, j \leq p} \left\{ \operatorname{Li}_2 \left( 1 - \frac{\zeta_i}{\lambda_j} \right) - \operatorname{Li}_2 \left( 1 - \frac{\eta_i}{\lambda_j} \right) \right\} - \sum_{i=1}^p \log^2((1 - c_1)\lambda_i) \right] \end{aligned}$$

→ Involves dilogarithm functions!

## Spectral clustering with feature $C_i$

### Setting:

- ▶ “ $m$ ” observations,  $X_1, \dots, X_m$  with  $X_i = [x_1^{(i)}, \dots, x_{n_i}^{(i)}]$
- ▶ **Two classes:**  $C_i = C^{(1)}$  for  $i \leq m/2$ ,  $C_i = C^{(2)}$  for  $i > m/2$ .

### Objective:

- ▶ Classify observations  $X_i$  based on  $C^{(1)}$  and  $C^{(2)}$ .

### Method:

- ▶ Spectral clustering with kernel

$$K_{ij} = D(C_i, C_j)$$

estimated by  $D(\hat{C}_i, \hat{C}_j)$  versus RMT estimator.

## Simulation: random $n_i$

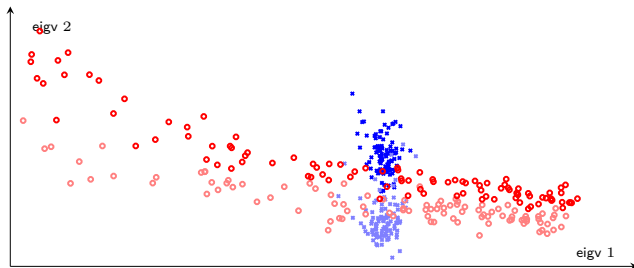


Figure: Eigenvectors 1 and 2 of  $K$  for traditional (red circles) versus RMT estimator (blue crosses).

### Classical

- ▶ Wide spread of eigenvectors
- ▶ Small inter space
- ▶ → Poor clustering

### RMT estimator

- ▶ Well centered eigenvector
- ▶ Large inter space
- ▶ → Good clustering



## Simulation: outlier $n_1 = \dots = n_{m-1}$ , $n_m = n_1/2$

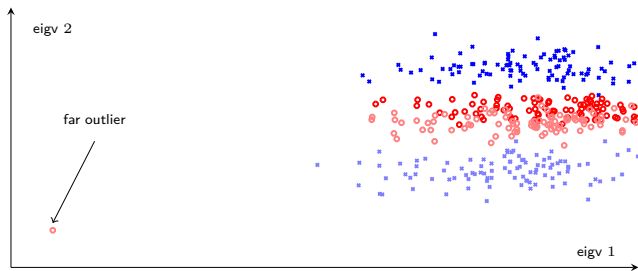


Figure: Eigenvectors 1 and 2 of  $K$  for traditional (red circles) versus RMT estimator (blue crosses).

### Classical

- ▶ Isolated outlier
- ▶ Adversarial effect of outlier ("draws" eigenvector to itself)
- ▶ Effect increased by more outliers

### RMT estimator

- ▶ No outlier effect
- ▶ Large inter space

### Observations:

- ▶  $X = [x_1, \dots, x_n]$ ,  $x_i \in \mathbb{R}^p$  with  $\mathbb{E}[x_i] = 0$ ,  $\mathbb{E}[x_i x_i^T] = C$ .

## Application to covariance matrix estimation

### Observations:

- ▶  $X = [x_1, \dots, x_n]$ ,  $x_i \in \mathbb{R}^p$  with  $\mathbb{E}[x_i] = 0$ ,  $\mathbb{E}[x_i x_i^T] = C$ .

### Objective:

- ▶ From the data  $x_i$ , estimate  $C$ .

## Application to covariance matrix estimation

### Observations:

- ▶  $X = [x_1, \dots, x_n]$ ,  $x_i \in \mathbb{R}^p$  with  $\mathbb{E}[x_i] = 0$ ,  $\mathbb{E}[x_i x_i^\top] = C$ .

### Objective:

- ▶ From the data  $x_i$ , estimate  $C$ .

### State of the Art:

- ▶ Sample Covariance Matrix (SCM):

$$\hat{C} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top = \frac{1}{n} X X^\top.$$

## Application to covariance matrix estimation

### Observations:

- ▶  $X = [x_1, \dots, x_n]$ ,  $x_i \in \mathbb{R}^p$  with  $\mathbb{E}[x_i] = 0$ ,  $\mathbb{E}[x_i x_i^\top] = C$ .

### Objective:

- ▶ From the data  $x_i$ , estimate  $C$ .

### State of the Art:

- ▶ Sample Covariance Matrix (SCM):

$$\hat{C} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top = \frac{1}{n} X X^\top.$$

→ Often justified by Law of Large Numbers:  $n \rightarrow \infty$ .

## Application to covariance matrix estimation

### Observations:

- ▶  $X = [x_1, \dots, x_n]$ ,  $x_i \in \mathbb{R}^p$  with  $\mathbb{E}[x_i] = 0$ ,  $\mathbb{E}[x_i x_i^\top] = C$ .

### Objective:

- ▶ From the data  $x_i$ , estimate  $C$ .

### State of the Art:

- ▶ Sample Covariance Matrix (SCM):

$$\hat{C} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top = \frac{1}{n} X X^\top.$$

→ Often justified by Law of Large Numbers:  $n \rightarrow \infty$ .

- ▶ Numerical inversion of asymptotic spectrum (QuEST).
  1. Bai-Silverstein equation: Estimate  $\lambda(\hat{C})$  from  $\lambda(C)$  in “large  $p, n$ ” regime.
  2. Need for non trivial inversion of the equation.

## Application to covariance matrix estimation (continued)

- ▶ Elementary idea

$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$

where  $\delta(M, C)$  can be the Fisher, Bhattacharyya, KL, Rényi divergence.

## Application to covariance matrix estimation (continued)

- ▶ Elementary idea

$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$

where  $\delta(M, C)$  can be the Fisher, Bhattacharyya, KL, Rényi divergence.

- ▶ Divergence  $\delta(M, C) = \int f(t) d\nu_p(t)$  inaccessible,  $\nu_p \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(M^{-1}C)}$ .



## Application to covariance matrix estimation (continued)

- ▶ Elementary idea

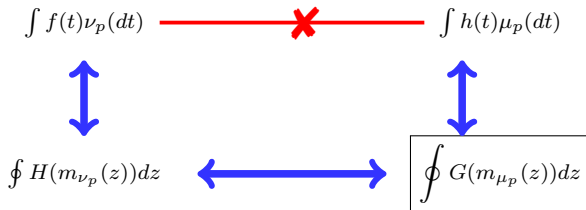
$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$

where  $\delta(M, C)$  can be the Fisher, Bhattacharyya, KL, Rényi divergence.

- ▶ Divergence  $\delta(M, C) = \int f(t) d\nu_p(t)$  inaccessible,  $\nu_p \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(M^{-1}C)}$ .

- ▶ Random Matrix improved estimate  $\hat{\delta}(M, X)$  of  $\delta(M, C)$  using

$$\mu_p \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(M^{-1}\hat{C})}$$



## Application to covariance matrix estimation (continued)

- ▶ Elementary idea

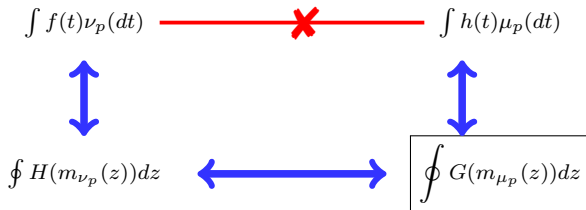
$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$

where  $\delta(M, C)$  can be the Fisher, Bhattacharyya, KL, Rényi divergence.

- ▶ Divergence  $\delta(M, C) = \int f(t) d\nu_p(t)$  inaccessible,  $\nu_p \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(M^{-1}C)}$ .

- ▶ Random Matrix improved estimate  $\hat{\delta}(M, X)$  of  $\delta(M, C)$  using

$$\mu_p \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(M^{-1}\hat{C})}$$



- ▶  $\hat{\delta}(M, X) < 0$  with non zero probability.

## Application to covariance matrix estimation (continued)

- ▶ Elementary idea

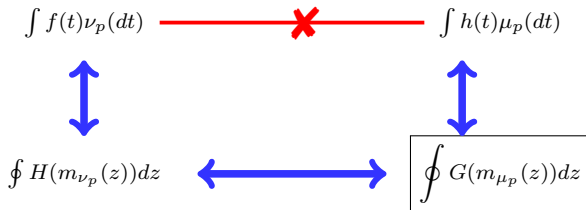
$$C \equiv \operatorname{argmin}_{M \succ 0} \delta(M, C)$$

where  $\delta(M, C)$  can be the Fisher, Bhattacharyya, KL, Rényi divergence.

- ▶ Divergence  $\delta(M, C) = \int f(t) d\nu_p(t)$  inaccessible,  $\nu_p \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(M^{-1}C)}$ .

- ▶ Random Matrix improved estimate  $\hat{\delta}(M, X)$  of  $\delta(M, C)$  using

$$\mu_p \equiv \frac{1}{p} \sum_{i=1}^p \delta_{\lambda_i(M^{-1}\hat{C})}$$



- ▶  $\hat{\delta}(M, X) < 0$  with non zero probability.
- ▶ RMT estimation

$$\check{C} \equiv \operatorname{argmin}_{M \succ 0} h(M), \quad h(M) = \hat{\delta}(M, X)^2$$

## Application to covariance matrix estimation (continued)

- ▶ Gradient descent over the Positive Definite manifold.

---

**Algorithm 1** RMT estimation algorithm.

---

**Require**  $M_0 \in C_n^{++}$ .

**Repeat**  $M \leftarrow M^{\frac{1}{2}} \exp\left(-tM^{-\frac{1}{2}} \nabla h_X(M) M^{-\frac{1}{2}}\right) M^{\frac{1}{2}}$  .

**Until** Convergence.

**Return**  $\check{C} = M$ .

---

## Application to covariance matrix estimation (continued)

- ▶ 2 Data classes  $x_1^{(1)}, \dots, x_{n_1}^{(1)} \sim N(\mu_1, C_1)$  and  $x_1^{(2)}, \dots, x_{n_2}^{(2)} \sim N(\mu_2, C_2)$ .
- ▶ Classify point  $x$  using Linear Discriminant Analysis based on the sign of

$$\delta_x^{\text{LDA}} = (\hat{\mu}_1 - \hat{\mu}_2)^\top \check{C}^{-1} x + \frac{1}{2} \hat{\mu}_2^\top \check{C}^{-1} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_1^\top \check{C}^{-1} \hat{\mu}_1.$$

- ▶ Estimate  $\check{C} \equiv \frac{n_1}{n_1+n_2} \check{C}_1 + \frac{n_2}{n_1+n_2} \check{C}_2$ .

## Application to covariance matrix estimation (continued)

- ▶ 2 Data classes  $x_1^{(1)}, \dots, x_{n_1}^{(1)} \sim N(\mu_1, C_1)$  and  $x_1^{(2)}, \dots, x_{n_2}^{(2)} \sim N(\mu_2, C_2)$ .
- ▶ Classify point  $x$  using Linear Discriminant Analysis based on the sign of

$$\delta_x^{\text{LDA}} = (\hat{\mu}_1 - \hat{\mu}_2)^T \check{C}^{-1} x + \frac{1}{2} \hat{\mu}_2^T \check{C}^{-1} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_1^T \check{C}^{-1} \hat{\mu}_1.$$

- ▶ Estimate  $\check{C} \equiv \frac{n_1}{n_1+n_2} \check{C}_1 + \frac{n_2}{n_1+n_2} \check{C}_2$ .

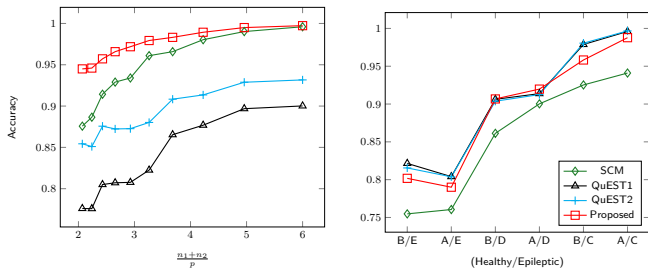


Figure: Mean accuracy obtained over 10 realizations of LDA classification. (Left)  $C_1$  and  $C_2$  Toeplitz-0.2/Toeplitz-0.4, and (Right) real EEG data.

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

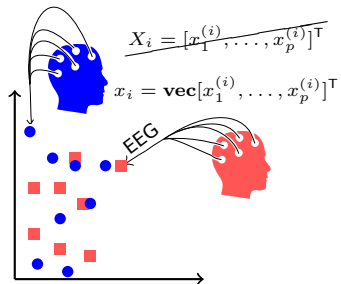
Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Reconsider clustering

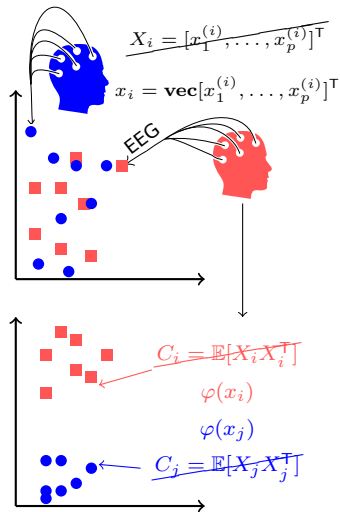
- ▶ Hard classification on raw data  $x_i$ :  
Need Features





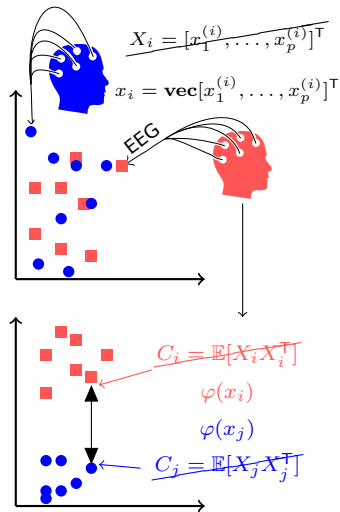
## Reconsider clustering

- ▶ Hard classification on raw data  $x_i$ :  
Need Features
- ▶ Relevant Feature: Covariance  $C_i$   
→ Learn features from data



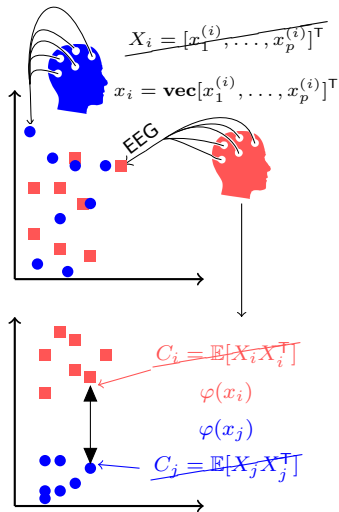
## Reconsider clustering

- ▶ Hard classification on raw data  $x_i$ :  
Need Features
- ▶ Relevant Feature: Covariance  $C_i$   
→ Learn features from data
- ▶  $D(C_i, C_j) \leftrightarrow \varphi(x_i)^T \varphi(x_j)$



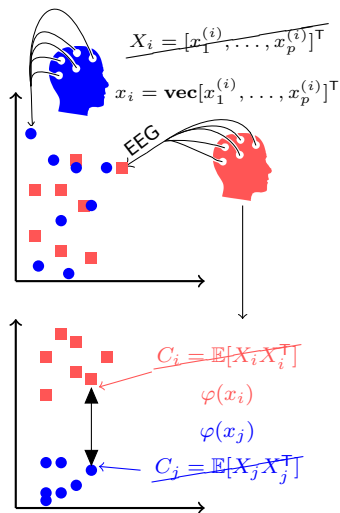
## Reconsider clustering

- ▶ Hard classification on raw data  $x_i$ :  
Need Features
- ▶ Relevant Feature: Covariance  $C_i$   
→ Learn features from data
- ▶  $D(C_i, C_j) \leftrightarrow \varphi(x_i)^T \varphi(x_j)$
- ▶ Kernel trick  
 $\varphi(x_i)^T \varphi(x_j) \rightarrow f(\|x_i - x_j\|^2)$  or  
 $f(x_i^T x_j)$



## Reconsider clustering

- ▶ Hard classification on raw data  $x_i$ :  
Need Features
- ▶ Relevant Feature: Covariance  $C_i$   
→ Learn features from data
- ▶  $D(C_i, C_j) \leftrightarrow \varphi(x_i)^T \varphi(x_j)$
- ▶ Kernel trick  
 $\varphi(x_i)^T \varphi(x_j) \rightarrow f(\|x_i - x_j\|^2)$  or  
 $f(x_i^T x_j)$
- ▶ Asymptotic performance of kernel methods?



## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

**Kernel Asymptotics**

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Problem Statement

- ▶ Dataset  $x_1, \dots, x_n \in \mathbb{R}^p$
- ▶ Objective: “cluster” data in  $k$  similarity classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ .

## Problem Statement

- ▶ Dataset  $x_1, \dots, x_n \in \mathbb{R}^p$
- ▶ Objective: “cluster” data in  $k$  similarity classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ .
- ▶ **Kernel spectral clustering** based on kernel matrix

$$K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$$

## Problem Statement

- ▶ Dataset  $x_1, \dots, x_n \in \mathbb{R}^p$
- ▶ Objective: “cluster” data in  $k$  similarity classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ .
- ▶ **Kernel spectral clustering** based on kernel matrix

$$K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$$

- ▶ Usually,  $\kappa(x, y) = f(x^T y)$  or  $\kappa(x, y) = f(\|x - y\|^2)$



# Kernel spectral clustering

Intuition (from small dimensions)

$$K = \begin{pmatrix} \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \gg 1 & \ll 1 & \ll 1 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \ll 1 & \gg 1 & \ll 1 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \ll 1 & \ll 1 & \gg 1 \\ \hline \end{array} \\ \hline \end{pmatrix} \begin{array}{l} \updownarrow C_1 \\ \updownarrow C_2 \\ \updownarrow C_3 \end{array}$$

- $K$  essentially low rank with class structure in eigenvectors.

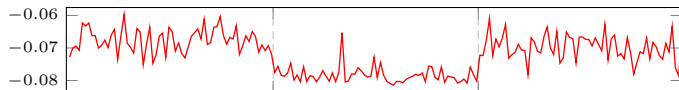
# Kernel spectral clustering

Intuition (from small dimensions)

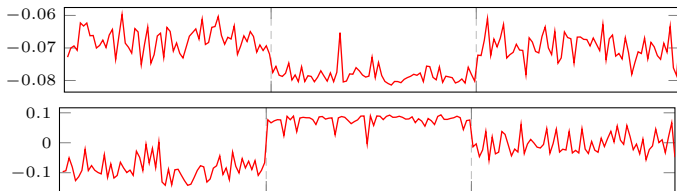
$$K = \begin{pmatrix} \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \gg 1 & \ll 1 & \ll 1 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \ll 1 & \gg 1 & \ll 1 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline \kappa(x_i, x_j) & \kappa(x_i, x_j) & \kappa(x_i, x_j) \\ \hline \ll 1 & \ll 1 & \gg 1 \\ \hline \end{array} \\ \hline \end{pmatrix} \begin{matrix} \updownarrow C_1 \\ \updownarrow C_2 \\ \updownarrow C_3 \end{matrix}$$

- ▶  $K$  essentially low rank with class structure in eigenvectors.
- ▶ Ng–Weiss–Jordan key remark:  $D^{-\frac{1}{2}} K D^{-\frac{1}{2}} (D^{\frac{1}{2}} j_a) \simeq D^{\frac{1}{2}} j_a$  ( $j_a$  canonical vector of  $C_a$ )

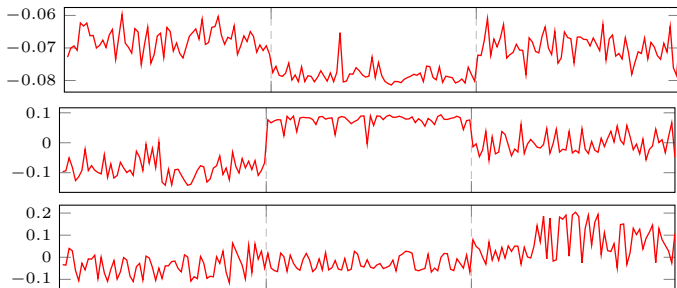
## Kernel Spectral Clustering



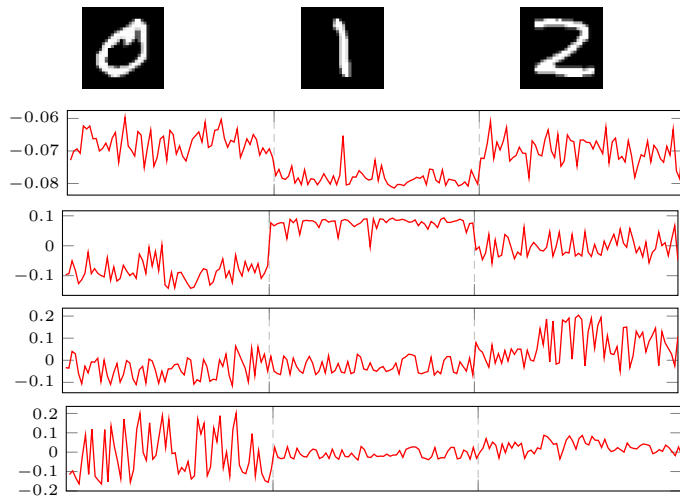
# Kernel Spectral Clustering



# Kernel Spectral Clustering



# Kernel Spectral Clustering



**Figure:** Leading four eigenvectors of  $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$  for MNIST data, RBF kernel ( $f(t) = \exp(-t^2/2)$ ).

# Kernel Spectral Clustering

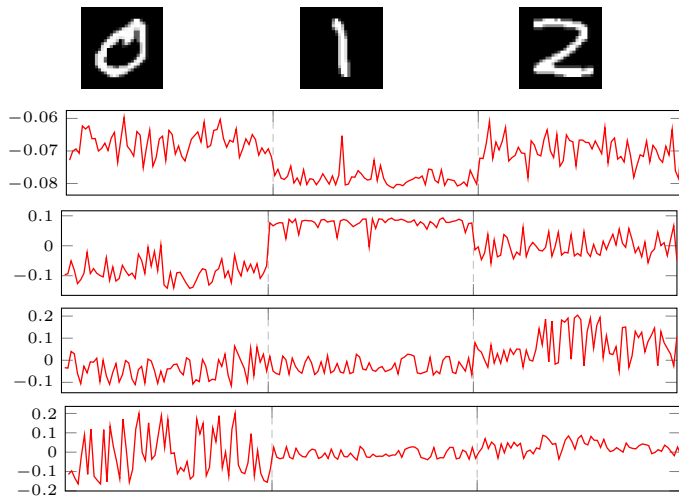


Figure: Leading four eigenvectors of  $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$  for MNIST data, RBF kernel ( $f(t) = \exp(-t^2/2)$ ).

- **Important Remark:** eigenvectors **informative** BUT far from  $D^{\frac{1}{2}} j_a$ !

# Model and Assumptions

## Gaussian mixture model:

- ▶  $x_1, \dots, x_n \in \mathbb{R}^p$ ,
- ▶  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ ,
- ▶  $x_1, \dots, x_{n_1} \in \mathcal{C}_1, \dots, x_{n-n_k+1}, \dots, x_n \in \mathcal{C}_k$ ,
- ▶  $x_i \sim \mathcal{N}(\mu_{g_i}, C_{g_i})$ .



# Model and Assumptions

## Gaussian mixture model:

- ▶  $x_1, \dots, x_n \in \mathbb{R}^p$ ,
- ▶  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ ,
- ▶  $x_1, \dots, x_{n_1} \in \mathcal{C}_1, \dots, x_{n-n_k+1}, \dots, x_n \in \mathcal{C}_k$ ,
- ▶  $x_i \sim \mathcal{N}(\mu_{g_i}, C_{g_i})$ .

## Assumption (Growth Rate)

As  $n \rightarrow \infty$ ,

1. **Data scaling:**  $\frac{p}{n} \rightarrow c_0 \in (0, \infty)$ ,  $\frac{n_a}{n} \rightarrow c_a \in (0, 1)$ ,
2. **Mean scaling:** with  $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$  and  $\mu_a^\circ \triangleq \mu_a - \mu^\circ$ , then  $\|\mu_a^\circ\| = O(1)$
3. **Covariance scaling:** with  $C^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} C_a$  and  $C_a^\circ \triangleq C_a - C^\circ$ , then

$$\|C_a\| = O(1), \quad \text{tr} C_a^\circ = O(\sqrt{p}), \quad \text{tr} C_a^\circ C_b^\circ = O(p)$$

# Model and Assumptions

## Gaussian mixture model:

- ▶  $x_1, \dots, x_n \in \mathbb{R}^p$ ,
- ▶  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ ,
- ▶  $x_1, \dots, x_{n_1} \in \mathcal{C}_1, \dots, x_{n-n_k+1}, \dots, x_n \in \mathcal{C}_k$ ,
- ▶  $x_i \sim \mathcal{N}(\mu_{g_i}, C_{g_i})$ .

## Assumption (Growth Rate)

As  $n \rightarrow \infty$ ,

1. **Data scaling:**  $\frac{p}{n} \rightarrow c_0 \in (0, \infty)$ ,  $\frac{n_a}{n} \rightarrow c_a \in (0, 1)$ ,
2. **Mean scaling:** with  $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$  and  $\mu_a^\circ \triangleq \mu_a - \mu^\circ$ , then  $\|\mu_a^\circ\| = O(1)$
3. **Covariance scaling:** with  $C^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} C_a$  and  $C_a^\circ \triangleq C_a - C^\circ$ , then

$$\|C_a\| = O(1), \quad \text{tr} C_a^\circ = O(\sqrt{p}), \quad \text{tr} C_a^\circ C_b^\circ = O(p)$$

For 2 classes, this is

$$\|\mu_1 - \mu_2\| = O(1), \quad \text{tr}(C_1 - C_2) = O(\sqrt{p}), \quad \|C_i\| = O(1), \quad \text{tr}[(C_1 - C_2)^2] = O(p).$$

# Model and Assumptions

## Gaussian mixture model:

- ▶  $x_1, \dots, x_n \in \mathbb{R}^p$ ,
- ▶  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ ,
- ▶  $x_1, \dots, x_{n_1} \in \mathcal{C}_1, \dots, x_{n-n_k+1}, \dots, x_n \in \mathcal{C}_k$ ,
- ▶  $x_i \sim \mathcal{N}(\mu_{g_i}, C_{g_i})$ .

## Assumption (Growth Rate)

As  $n \rightarrow \infty$ ,

1. **Data scaling:**  $\frac{p}{n} \rightarrow c_0 \in (0, \infty)$ ,  $\frac{n_a}{n} \rightarrow c_a \in (0, 1)$ ,
2. **Mean scaling:** with  $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$  and  $\mu_a^\circ \triangleq \mu_a - \mu^\circ$ , then  $\|\mu_a^\circ\| = O(1)$
3. **Covariance scaling:** with  $C^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} C_a$  and  $C_a^\circ \triangleq C_a - C^\circ$ , then

$$\|C_a\| = O(1), \quad \text{tr} C_a^\circ = O(\sqrt{p}), \quad \text{tr} C_a^\circ C_b^\circ = O(p)$$

For 2 classes, this is

$$\|\mu_1 - \mu_2\| = O(1), \quad \text{tr}(C_1 - C_2) = O(\sqrt{p}), \quad \|C_i\| = O(1), \quad \text{tr}([C_1 - C_2]^2) = O(p).$$

## Remark: [Neyman–Pearson optimality]

- ▶  $x \sim \mathcal{N}(\pm\mu, I_p)$  (known  $\mu$ ) decidable iff  $\|\mu\| \geq O(1)$ .
- ▶  $x \sim \mathcal{N}(0, (1 \pm \varepsilon)I_p)$  (known  $\varepsilon$ ) decidable iff  $\|\varepsilon\| \geq O(p^{-\frac{1}{2}})$ .

### Kernel Matrix:

- ▶ Kernel matrix of interest:

$$K = \left\{ f \left( \frac{1}{p} \|x_i - x_j\|^2 \right) \right\}_{i,j=1}^n$$

for some sufficiently smooth nonnegative  $f$  ( $f(\frac{1}{p} x_i^\top x_j)$  simpler).

### Kernel Matrix:

- ▶ Kernel matrix of interest:

$$K = \left\{ f \left( \frac{1}{p} \|x_i - x_j\|^2 \right) \right\}_{i,j=1}^n$$

for some sufficiently smooth nonnegative  $f$  ( $f(\frac{1}{p}x_i^\top x_j)$  simpler).

- ▶ We study the normalized Laplacian:

$$L = nD^{-\frac{1}{2}} \left( K - \frac{dd^\top}{d^\top 1_n} \right) D^{-\frac{1}{2}}$$

with  $d = K1_n$ ,  $D = \text{diag}(d)$ .

*(more stable both theoretically and in practice)*

- **Key Remark:** Under growth rate assumptions,

$$\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \|x_i - x_j\|^2 - \tau \right| \right\} \xrightarrow{\text{a.s.}} 0.$$

where  $\tau = \frac{1}{p} \text{tr} C^\circ$ .

- **Key Remark:** Under growth rate assumptions,

$$\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \|x_i - x_j\|^2 - \tau \right| \right\} \xrightarrow{\text{a.s.}} 0.$$

where  $\tau = \frac{1}{p} \text{tr } C^\circ$ .

⇒ Suggests that (up to diagonal)  $K \simeq f(\tau) 1_n 1_n^\top$ !

- ▶ **Key Remark:** Under growth rate assumptions,

$$\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \|x_i - x_j\|^2 - \tau \right| \right\} \xrightarrow{\text{a.s.}} 0.$$

where  $\tau = \frac{1}{p} \text{tr} C^\circ$ .

⇒ Suggests that (up to diagonal)  $K \simeq f(\tau) 1_n 1_n^\top$ !

- ▶ In fact, **information hidden in low order fluctuations!** from “matrix-wise” Taylor expansion of  $K$ :

$$K = \underbrace{f(\tau) 1_n 1_n^\top}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n} K_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{K_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$



- ▶ **Key Remark:** Under growth rate assumptions,

$$\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \|x_i - x_j\|^2 - \tau \right| \right\} \xrightarrow{\text{a.s.}} 0.$$

where  $\tau = \frac{1}{p} \text{tr} C^\circ$ .

⇒ Suggests that (up to diagonal)  $K \simeq f(\tau)1_n 1_n^T$ !

- ▶ In fact, **information hidden in low order fluctuations!** from “matrix-wise” Taylor expansion of  $K$ :

$$K = \underbrace{f(\tau)1_n 1_n^T}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}K_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{K_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

Clearly not the (small dimension) expected behavior.

## Random Matrix Equivalent

Theorem (Random Matrix Equivalent [Couillet, Benaych'2015])

As  $n, p \rightarrow \infty$ ,  $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$ , where

$$L = nD^{-\frac{1}{2}} \left( K - \frac{dd^T}{d^T 1_n} \right) D^{-\frac{1}{2}}, \text{ avec } K_{ij} = f \left( \frac{1}{p} \|x_i - x_j\|^2 \right)$$

$$\hat{L} = -2 \frac{f'(\tau)}{f(\tau)} \frac{1}{p} PW^T WP + \frac{1}{p} JBJ^T + *$$

et  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} 1_n 1_n^T$ ,

## Random Matrix Equivalent

Theorem (Random Matrix Equivalent [Couillet, Benaych'2015])

As  $n, p \rightarrow \infty$ ,  $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$ , where

$$L = nD^{-\frac{1}{2}} \left( K - \frac{dd^T}{d^T 1_n} \right) D^{-\frac{1}{2}}, \text{ avec } K_{ij} = f \left( \frac{1}{p} \|x_i - x_j\|^2 \right)$$

$$\hat{L} = -2 \frac{f'(\tau)}{f(\tau)} \frac{1}{p} P W^T W P + \frac{1}{p} J B J^T + *$$

et  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} 1_n 1_n^T$ ,

$$J = [j_1, \dots, j_k], j_a^T = (0 \dots 0, 1_{n_a}, 0, \dots, 0)$$

$$B = -2 \frac{f'(\tau)}{f(\tau)} M^T M + \left( \frac{f''(\tau)}{f(\tau)} - \frac{5f'(\tau)^2}{4f(\tau)^2} \right) t t^T + 2 \frac{f''(\tau)}{f(\tau)} T + *.$$

Recall  $M = [\mu_1^\circ, \dots, \mu_k^\circ]$ ,  $t = [\frac{1}{\sqrt{p}} \text{tr} C_1^\circ, \dots, \frac{1}{\sqrt{p}} \text{tr} C_k^\circ]^T$ ,  $T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k$ .

## Random Matrix Equivalent

### Theorem (Random Matrix Equivalent [Couillet, Benaych'2015])

As  $n, p \rightarrow \infty$ ,  $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$ , where

$$L = nD^{-\frac{1}{2}} \left( K - \frac{dd^T}{d^T 1_n} \right) D^{-\frac{1}{2}}, \text{ avec } K_{ij} = f \left( \frac{1}{p} \|x_i - x_j\|^2 \right)$$

$$\hat{L} = -2 \frac{f'(\tau)}{f(\tau)} \frac{1}{p} P W^T W P + \frac{1}{p} J B J^T + *$$

et  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} 1_n 1_n^T$ ,

$$J = [j_1, \dots, j_k], j_a^T = (0 \dots 0, 1_{n_a}, 0, \dots, 0)$$

$$B = -2 \frac{f'(\tau)}{f(\tau)} M^T M + \left( \frac{f''(\tau)}{f(\tau)} - \frac{5f'(\tau)^2}{4f(\tau)^2} \right) t t^T + 2 \frac{f''(\tau)}{f(\tau)} T + *.$$

Recall  $M = [\mu_1^\circ, \dots, \mu_k^\circ]$ ,  $t = [\frac{1}{\sqrt{p}} \text{tr} C_1^\circ, \dots, \frac{1}{\sqrt{p}} \text{tr} C_k^\circ]^T$ ,  $T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k$ .

**Fundamental conclusions:**

## Random Matrix Equivalent

### Theorem (Random Matrix Equivalent [Couillet, Benaych'2015])

As  $n, p \rightarrow \infty$ ,  $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$ , where

$$L = nD^{-\frac{1}{2}} \left( K - \frac{dd^T}{d^T 1_n} \right) D^{-\frac{1}{2}}, \text{ avec } K_{ij} = f \left( \frac{1}{p} \|x_i - x_j\|^2 \right)$$
$$\hat{L} = -2 \frac{f'(\tau)}{f(\tau)} \frac{1}{p} P W^T W P + \frac{1}{p} J B J^T + *$$

et  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} 1_n 1_n^T$ ,

$$J = [j_1, \dots, j_k], j_a^T = (0 \dots 0, 1_{n_a}, 0, \dots, 0)$$

$$B = -2 \frac{f'(\tau)}{f(\tau)} M^T M + \left( \frac{f''(\tau)}{f(\tau)} - \frac{5f'(\tau)^2}{4f(\tau)^2} \right) t t^T + 2 \frac{f''(\tau)}{f(\tau)} T + *.$$

Recall  $M = [\mu_1^\circ, \dots, \mu_k^\circ]$ ,  $t = [\frac{1}{\sqrt{p}} \text{tr} C_1^\circ, \dots, \frac{1}{\sqrt{p}} \text{tr} C_k^\circ]^T$ ,  $T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k$ .

#### Fundamental conclusions:

- ▶ asymptotic kernel impact only through  $f'(\tau)$  and  $f''(\tau)$ , that's all!

## Random Matrix Equivalent

### Theorem (Random Matrix Equivalent [Couillet, Benaych'2015])

As  $n, p \rightarrow \infty$ ,  $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$ , where

$$L = nD^{-\frac{1}{2}} \left( K - \frac{dd^T}{d^T 1_n} \right) D^{-\frac{1}{2}}, \text{ avec } K_{ij} = f \left( \frac{1}{p} \|x_i - x_j\|^2 \right)$$

$$\hat{L} = -2 \frac{f'(\tau)}{f(\tau)} \frac{1}{p} P W^T W P + \frac{1}{p} J B J^T + *$$

et  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} 1_n 1_n^T$ ,

$$J = [j_1, \dots, j_k], j_a^T = (0 \dots 0, 1_{n_a}, 0, \dots, 0)$$

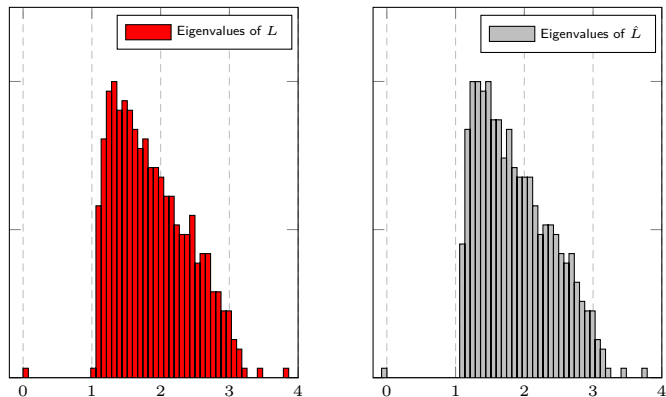
$$B = -2 \frac{f'(\tau)}{f(\tau)} M^T M + \left( \frac{f''(\tau)}{f(\tau)} - \frac{5f'(\tau)^2}{4f(\tau)^2} \right) t t^T + 2 \frac{f''(\tau)}{f(\tau)} T + *.$$

Recall  $M = [\mu_1^\circ, \dots, \mu_k^\circ]$ ,  $t = [\frac{1}{\sqrt{p}} \text{tr} C_1^\circ, \dots, \frac{1}{\sqrt{p}} \text{tr} C_k^\circ]^T$ ,  $T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k$ .

#### Fundamental conclusions:

- ▶ asymptotic kernel impact only through  $f'(\tau)$  and  $f''(\tau)$ , that's all!
- ▶ spectral clustering reads  $M^T M$ ,  $t t^T$  and  $T$ , that's all!

## Isolated eigenvalues: Gaussian inputs



**Figure:** Eigenvalues of  $L$  and  $\hat{L}$ ,  $k = 3$ ,  $p = 2048$ ,  $n = 512$ ,  $c_1 = c_2 = 1/4$ ,  $c_3 = 1/2$ ,  $[\mu_a]_j = 4\delta_{aj}$ ,  $C_a = (1 + 2(a - 1)/\sqrt{p})I_p$ ,  $f(x) = \exp(-x/2)$ .

## Theoretical Findings versus MNIST

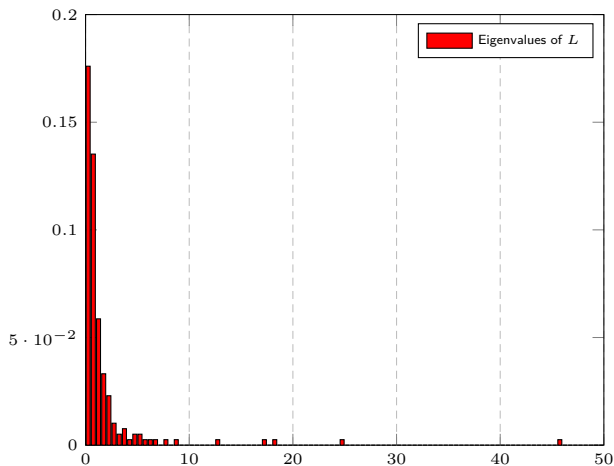


Figure: Eigenvalues of  $L$  (red) and (equivalent Gaussian model)  $\hat{L}$  (white), MNIST data,  $p = 784$ ,  $n = 192$ .



# Theoretical Findings versus MNIST

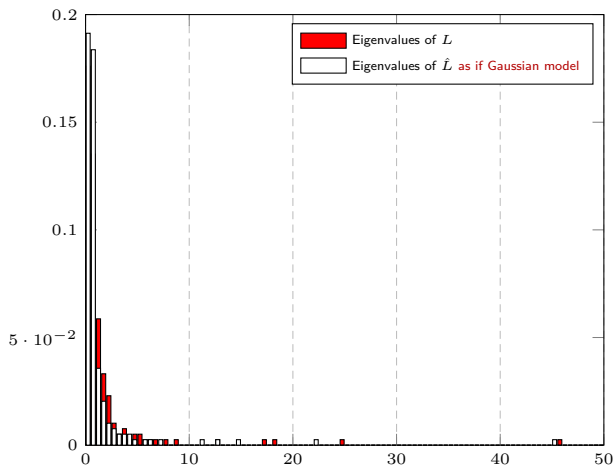


Figure: Eigenvalues of  $L$  (red) and (equivalent Gaussian model)  $\hat{L}$  (white), MNIST data,  $p = 784$ ,  $n = 192$ .

# Theoretical Findings versus MNIST

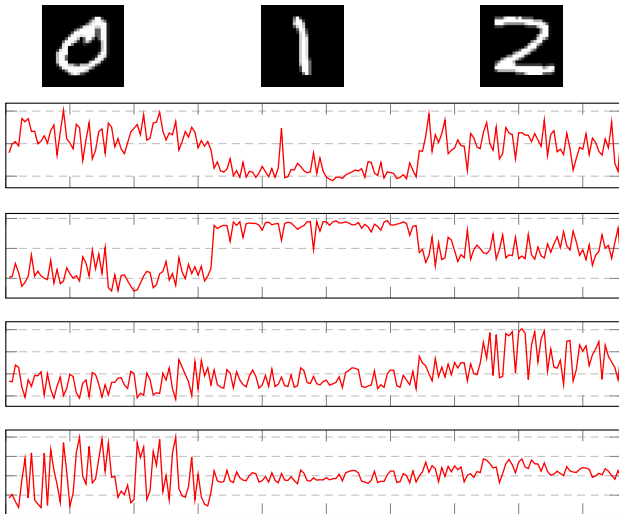


Figure: Leading four eigenvectors of  $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$  for MNIST data (red) and theoretical findings (blue).

# Theoretical Findings versus MNIST

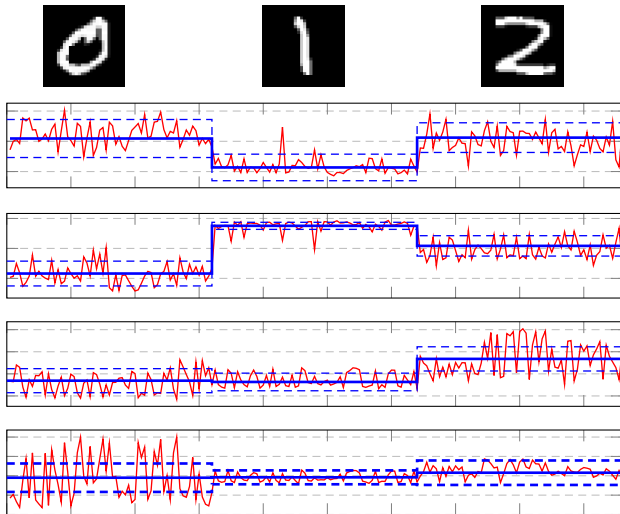
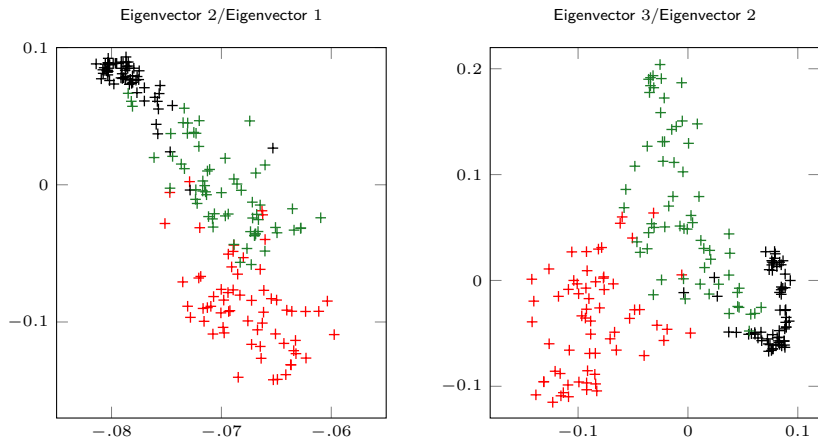


Figure: Leading four eigenvectors of  $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$  for MNIST data (red) and theoretical findings (blue).

# Theoretical Findings versus MNIST



**Figure:** 2D representation of eigenvectors of  $L$ , for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

# Theoretical Findings versus MNIST

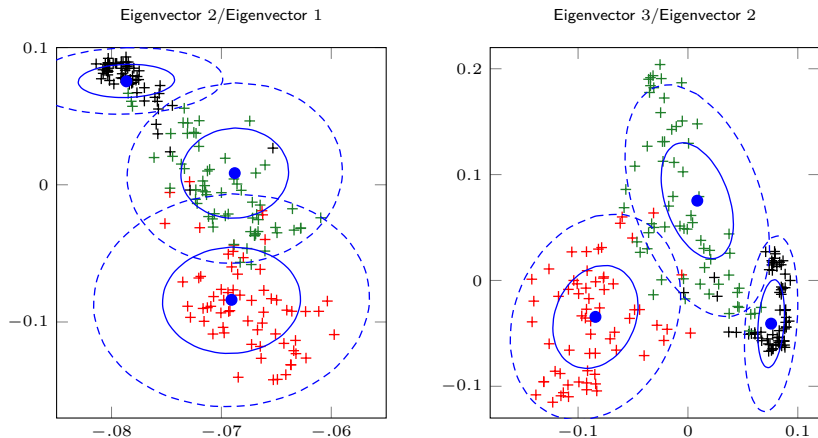


Figure: 2D representation of eigenvectors of  $L$ , for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in blue. Class 1 in red, Class 2 in black, Class 3 in green.

## The surprising $f'(\tau) = 0$ case

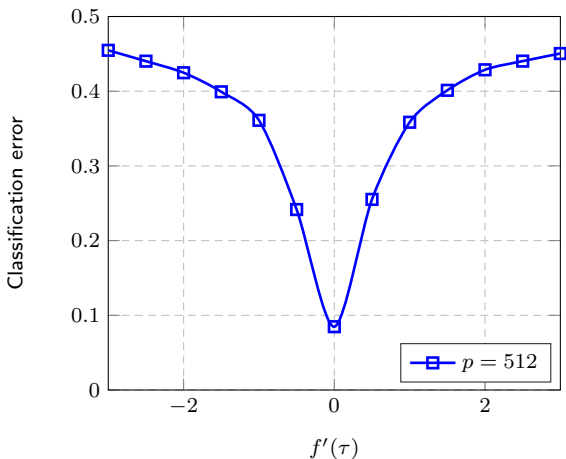


Figure: Polynomial kernel with  $f(\tau) = 4$ ,  $f''(\tau) = 2$ ,  $x_i \in \mathcal{N}(0, C_a)$ , with  $C_1 = I_p$ ,  $[C_2]_{i,j} = .4^{|i-j|}$ ,  $c_0 = \frac{1}{4}$ .

## The surprising $f'(\tau) = 0$ case

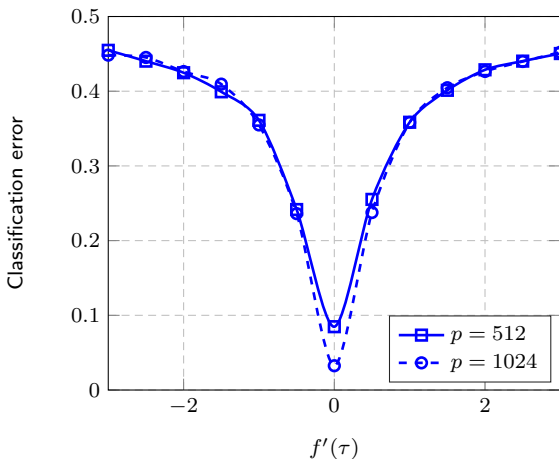


Figure: Polynomial kernel with  $f(\tau) = 4$ ,  $f''(\tau) = 2$ ,  $x_i \in \mathcal{N}(0, C_a)$ , with  $C_1 = I_p$ ,  $[C_2]_{i,j} = .4^{|i-j|}$ ,  $c_0 = \frac{1}{4}$ .

## The surprising $f'(\tau) = 0$ case

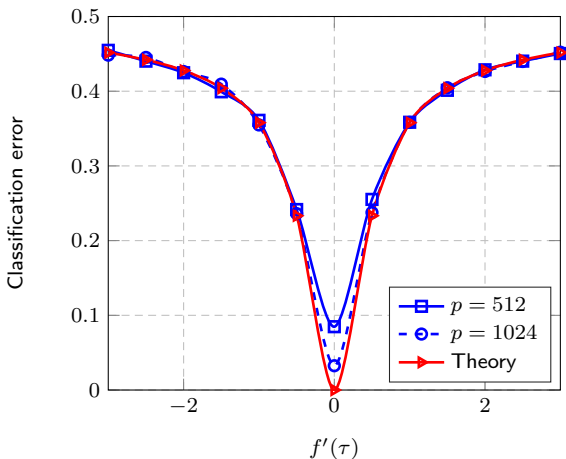


Figure: Polynomial kernel with  $f(\tau) = 4$ ,  $f''(\tau) = 2$ ,  $x_i \in \mathcal{N}(0, C_a)$ , with  $C_1 = I_p$ ,  $[C_2]_{i,j} = .4^{|i-j|}$ ,  $c_0 = \frac{1}{4}$ .



## The surprising $f'(\tau) = 0$ case

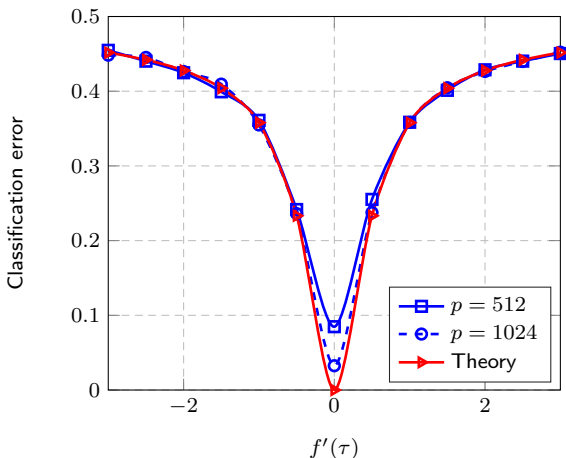


Figure: Polynomial kernel with  $f(\tau) = 4$ ,  $f''(\tau) = 2$ ,  $x_i \in \mathcal{N}(0, C_a)$ , with  $C_1 = I_p$ ,  $[C_2]_{i,j} = .4^{|i-j|}$ ,  $c_0 = \frac{1}{4}$ .

- **Trivial classification** when  $t = 0$ ,  $M = 0$  and  $\|T\| = O(1)$ .

# Spectral Clustering: The case $f'(\tau) = 0$

Position of the problem

**Problem:** Cluster large data  $x_1, \dots, x_n \in \mathbb{R}^p$  based on “spanned subspaces”.

# Spectral Clustering: The case $f'(\tau) = 0$

Position of the problem

**Problem:** Cluster large data  $x_1, \dots, x_n \in \mathbb{R}^p$  based on “spanned subspaces”.

**Method:**

- ▶ Still assume  $x_1, \dots, x_n$  belong to  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ .
- ▶ Zero-mean Gaussian model for the data: for  $x_i \in \mathcal{C}_k$ ,

$$x_i \sim \mathcal{N}(0, C_k).$$

# Spectral Clustering: The case $f'(\tau) = 0$

Position of the problem

**Problem:** Cluster large data  $x_1, \dots, x_n \in \mathbb{R}^p$  based on “spanned subspaces”.

**Method:**

- ▶ Still assume  $x_1, \dots, x_n$  belong to  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ .
- ▶ Zero-mean Gaussian model for the data: for  $x_i \in \mathcal{C}_k$ ,

$$x_i \sim \mathcal{N}(0, C_k).$$

- ▶ Performance of  $L = nD^{-\frac{1}{2}} \left( K - \frac{1_n 1_n^T}{1_n^T D 1_n} \right) D^{-\frac{1}{2}}$ , with

$$K = \left\{ f \left( \|\bar{x}_i - \bar{x}_j\|^2 \right) \right\}_{1 \leq i, j \leq n}, \quad \bar{x} = \frac{x}{\|x\|}$$

in the regime  $n, p \rightarrow \infty$ .

(alternatively, we can ask  $\frac{1}{p} \text{tr} C_i = 1$  for all  $1 \leq i \leq k$ )

# Spectral Clustering: The case $f'(\tau) = 0$

Model and Reminders

**Assumption 1 [Classes].** Vectors  $x_1, \dots, x_n \in \mathbb{R}^p$  i.i.d. from  $k$ -class Gaussian mixture, with  $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$  (sorted by class for simplicity).

# Spectral Clustering: The case $f'(\tau) = 0$

## Model and Reminders

**Assumption 1 [Classes].** Vectors  $x_1, \dots, x_n \in \mathbb{R}^p$  i.i.d. from  $k$ -class Gaussian mixture, with  $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$  (sorted by class for simplicity).

**Assumption 2a [Growth Rates].** As  $n \rightarrow \infty$ , for each  $a \in \{1, \dots, k\}$ ,

1.  $\frac{n}{p} \rightarrow c_0 \in (0, \infty)$
2.  $\frac{n_a}{n} \rightarrow c_a \in (0, \infty)$
3.  $\frac{1}{p} \text{tr} C_a = 1$  and  $\text{tr} C_a^\circ C_b^\circ = O(p)$ , with  $C_a^\circ = C_a - C^\circ$ ,  $C^\circ = \sum_{b=1}^k c_b C_b$ .

# Spectral Clustering: The case $f'(\tau) = 0$

Model and Reminders

**Assumption 1 [Classes].** Vectors  $x_1, \dots, x_n \in \mathbb{R}^p$  i.i.d. from  $k$ -class Gaussian mixture, with  $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$  (sorted by class for simplicity).

**Assumption 2a [Growth Rates].** As  $n \rightarrow \infty$ , for each  $a \in \{1, \dots, k\}$ ,

1.  $\frac{n}{p} \rightarrow c_0 \in (0, \infty)$
2.  $\frac{n_a}{n} \rightarrow c_a \in (0, \infty)$
3.  $\frac{1}{p} \text{tr} C_a = 1$  and  $\text{tr} C_a^\circ C_b^\circ = O(p)$ , with  $C_a^\circ = C_a - C^\circ$ ,  $C^\circ = \sum_{b=1}^k c_b C_b$ .

## Theorem (Corollary of Previous Section)

Let  $f$  smooth with  $f'(2) \neq 0$ . Then, under Assumptions 2a,

$$L = nD^{-\frac{1}{2}} \left( K - \frac{1_n 1_n^\top}{1_n^\top D 1_n} \right) D^{-\frac{1}{2}}, \text{ with } K = \left\{ f(\|\bar{x}_i - \bar{x}_j\|^2) \right\}_{i,j=1}^n \quad (\bar{x} = x/\|x\|)$$

exhibits *phase transition phenomenon*

# Spectral Clustering: The case $f'(\tau) = 0$

## Model and Reminders

**Assumption 1 [Classes].** Vectors  $x_1, \dots, x_n \in \mathbb{R}^p$  i.i.d. from  $k$ -class Gaussian mixture, with  $x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(0, C_k)$  (sorted by class for simplicity).

**Assumption 2a [Growth Rates].** As  $n \rightarrow \infty$ , for each  $a \in \{1, \dots, k\}$ ,

1.  $\frac{n}{p} \rightarrow c_0 \in (0, \infty)$
2.  $\frac{n_a}{n} \rightarrow c_a \in (0, \infty)$
3.  $\frac{1}{p} \text{tr} C_a = 1$  and  $\text{tr} C_a^\circ C_b^\circ = O(p)$ , with  $C_a^\circ = C_a - C^\circ$ ,  $C^\circ = \sum_{b=1}^k c_b C_b$ .

## Theorem (Corollary of Previous Section)

Let  $f$  smooth with  $f'(2) \neq 0$ . Then, under Assumptions 2a,

$$L = nD^{-\frac{1}{2}} \left( K - \frac{1_n 1_n^\top}{1_n^\top D 1_n} \right) D^{-\frac{1}{2}}, \text{ with } K = \left\{ f(\|\bar{x}_i - \bar{x}_j\|^2) \right\}_{i,j=1}^n \quad (\bar{x} = x/\|x\|)$$

exhibits **phase transition phenomenon**, i.e., leading eigenvectors of  $L$  asymptotically contain structural information about  $\mathcal{C}_1, \dots, \mathcal{C}_k$  **if and only if**

$$T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k$$

has sufficiently large eigenvalues (here  $M = 0$ ,  $t = 0$ ).



# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

**Assumption 2b [Growth Rates].** As  $n \rightarrow \infty$ , for each  $a \in \{1, \dots, k\}$ ,

1.  $\frac{n}{p} \rightarrow c_0 \in (0, \infty)$
2.  $\frac{n_a}{n} \rightarrow c_a \in (0, \infty)$
3.  $\frac{1}{p} \text{tr} C_a = 1$  and  $\text{tr} C_a^\circ C_b^\circ = O(p)$ , with  $C_a^\circ = C_a - C^\circ$ ,  $C^\circ = \sum_{b=1}^k c_b C_b$ .

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

**Assumption 2b [Growth Rates].** As  $n \rightarrow \infty$ , for each  $a \in \{1, \dots, k\}$ ,

1.  $\frac{n}{p} \rightarrow c_0 \in (0, \infty)$
2.  $\frac{n_a}{n} \rightarrow c_a \in (0, \infty)$
3.  $\frac{1}{p} \text{tr} C_a = 1$  and  $\text{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$ , with  $C_a^\circ = C_a - C^\circ$ ,  $C^\circ = \sum_{b=1}^k c_b C_b$ .

(in this regime, *previous kernels clearly fail*)

**Remark: [Neyman–Pearson optimality]**

- if  $C_i = I_p \pm E$  with  $\|E\| \rightarrow 0$ , **detectability** iff  $\frac{1}{p} \text{tr}(C_1 - C_2)^2 \geq O(p^{-\frac{1}{2}})$ .

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

**Assumption 2b [Growth Rates].** As  $n \rightarrow \infty$ , for each  $a \in \{1, \dots, k\}$ ,

1.  $\frac{n}{p} \rightarrow c_0 \in (0, \infty)$
2.  $\frac{n_a}{n} \rightarrow c_a \in (0, \infty)$
3.  $\frac{1}{p} \text{tr} C_a = 1$  and  $\text{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$ , with  $C_a^\circ = C_a - C^\circ$ ,  $C^\circ = \sum_{b=1}^k c_b C_b$ .

(in this regime, *previous kernels clearly fail*)

**Remark: [Neyman–Pearson optimality]**

- if  $C_i = I_p \pm E$  with  $\|E\| \rightarrow 0$ , **detectability** iff  $\frac{1}{p} \text{tr}(C_1 - C_2)^2 \geq O(p^{-\frac{1}{2}})$ .

## Theorem (Random Equivalent for $f'(2) = 0$ )

Let  $f$  be smooth with  $f'(2) = 0$  and

$$\mathcal{L} \equiv \sqrt{p} \frac{f(2)}{2f''(2)} \left[ L - \frac{f(0) - f(2)}{f(2)} P \right], \quad P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

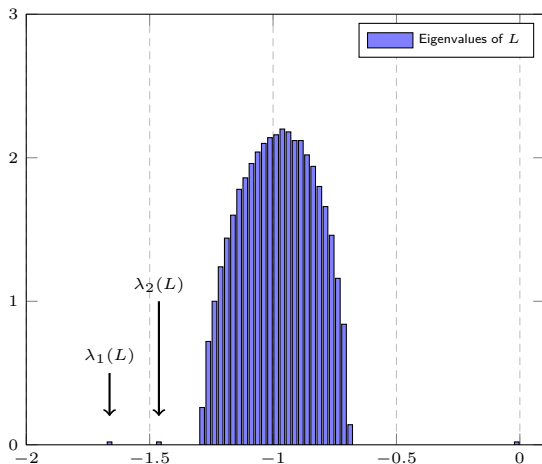
Then, under Assumptions 2b,

$$\mathcal{L} = P \Phi P + \left\{ \frac{1}{\sqrt{p}} \text{tr}(C_a^\circ C_b^\circ) \frac{\mathbf{1}_{n_a} \mathbf{1}_{n_b}^\top}{p} \right\}_{a,b=1}^k + o_{\|\cdot\|}(1)$$

where  $\Phi_{ij} = \delta_{i \neq j} \sqrt{p} \left[ (x_i^\top x_j)^2 - E[(x_i^\top x_j)^2] \right]$ .

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

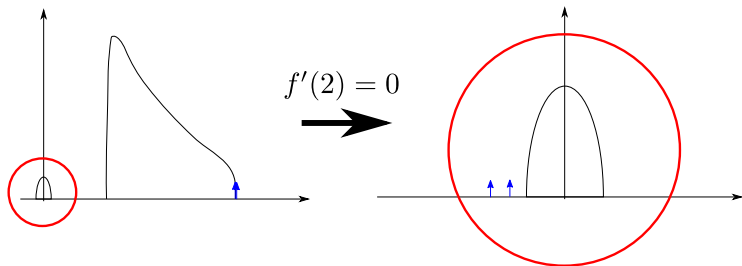


**Figure:** Eigenvalues of  $L$ ,  $p = 1000$ ,  $n = 2000$ ,  $k = 3$ ,  $c_1 = c_2 = 1/4$ ,  $c_3 = 1/2$ ,  
 $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^T$ ,  $W_i \in \mathbb{R}^{p \times (p/8)}$  of i.i.d.  $\mathcal{N}(0, 1)$  entries,  $f(t) = \exp(-(t - 2)^2)$ .

**$\Rightarrow$  No longer a Marcenko–Pastur like bulk, but rather a semi-circle bulk!**

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$



# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

**Roadmap.** We now need to:

- ▶ study the spectrum of  $\Phi$

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

**Roadmap.** We now need to:

- ▶ study the spectrum of  $\Phi$
- ▶ study the isolated eigenvalues of  $\mathcal{L}$  (and the phase transition)

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

**Roadmap.** We now need to:

- ▶ study the spectrum of  $\Phi$
- ▶ study the isolated eigenvalues of  $\mathcal{L}$  (and the phase transition)
- ▶ retrieve information from the eigenvectors.



# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

**Roadmap.** We now need to:

- ▶ study the spectrum of  $\Phi$
- ▶ study the isolated eigenvalues of  $\mathcal{L}$  (and the phase transition)
- ▶ retrieve information from the eigenvectors.

## Theorem (Semi-circle law for $\Phi$ )

Let  $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(\mathcal{L})}$ . Then, under Assumption 2b,

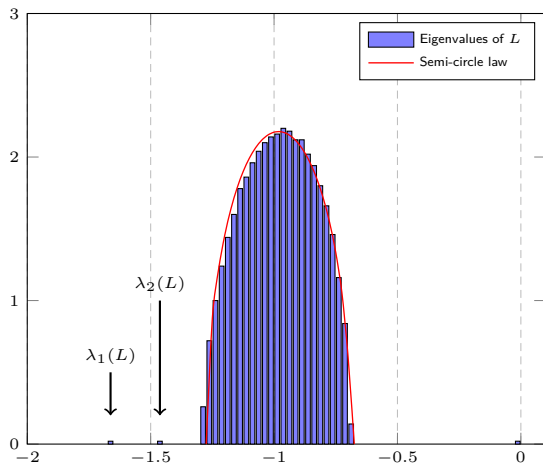
$$\mu_n \xrightarrow{\text{a.s.}} \mu$$

with  $\mu$  the semi-circle distribution

$$\mu(dt) = \frac{1}{2\pi c_0 \omega^2} \sqrt{(4c_0 \omega^2 - t^2)^+} dt, \quad \omega = \lim_{p \rightarrow \infty} \sqrt{2} \frac{1}{p} \text{tr}(C^\circ)^2.$$

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$



**Figure:** Eigenvalues of  $L$ ,  $p = 1000$ ,  $n = 2000$ ,  $k = 3$ ,  $c_1 = c_2 = 1/4$ ,  $c_3 = 1/2$ ,  $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^T$ ,  $W_i \in \mathbb{R}^{p \times (p/8)}$  of i.i.d.  $\mathcal{N}(0, 1)$  entries,  $f(t) = \exp(-(t - 2)^2)$ .

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

Denote now

$$\mathcal{T} \equiv \lim_{p \rightarrow \infty} \left\{ \frac{\sqrt{c_a c_b}}{\sqrt{p}} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k .$$

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

Denote now

$$\mathcal{T} \equiv \lim_{p \rightarrow \infty} \left\{ \frac{\sqrt{c_a c_b}}{\sqrt{p}} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k.$$

## Theorem (Isolated Eigenvalues)

Let  $\nu_1 \geq \dots \geq \nu_k$  eigenvalues of  $\mathcal{T}$ . Then, if  $\sqrt{c_0} |\nu_i| > \omega$ ,  $\mathcal{L}$  has an isolated eigenvalue  $\lambda_i$  satisfying

$$\lambda_i \xrightarrow{\text{a.s.}} \rho_i \equiv c_0 \nu_i + \frac{\omega^2}{\nu_i}.$$

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

## Theorem (Isolated Eigenvectors)

For each isolated eigenpair  $(\lambda_i, u_i)$  of  $\mathcal{L}$  corresponding to  $(\nu_i, v_i)$  of  $\mathcal{T}$ , write

$$u_i = \sum_{a=1}^k \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

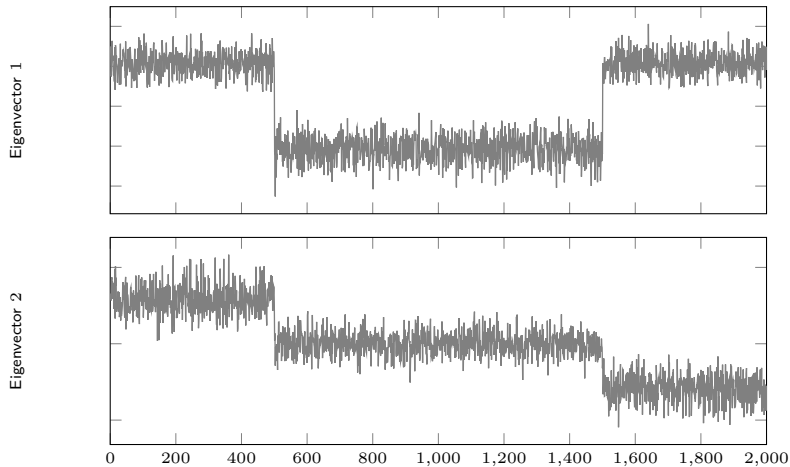
with  $j_a = [0_{n_1}^\top, \dots, 1_{n_a}^\top, \dots, 0_{n_k}^\top]^\top$ ,  $(w_i^a)^\top j_a = 0$ ,  $\text{supp}(w_i^a) = \text{supp}(j_a)$ ,  $\|w_i^a\| = 1$ .  
Then, under Assumptions 1–2b,

$$\alpha_i^a \alpha_i^b \xrightarrow{\text{a.s.}} \left(1 - \frac{1}{c_0} \frac{\omega^2}{\nu_i^2}\right) [v_i v_i^\top]_{ab}$$
$$(\sigma_i^a)^2 \xrightarrow{\text{a.s.}} \frac{c_a}{c_0} \frac{\omega^2}{\nu_i^2}$$

and the fluctuations of  $u_i, u_j$ ,  $i \neq j$ , are asymptotically uncorrelated.

# Spectral Clustering: The case $f'(\tau) = 0$

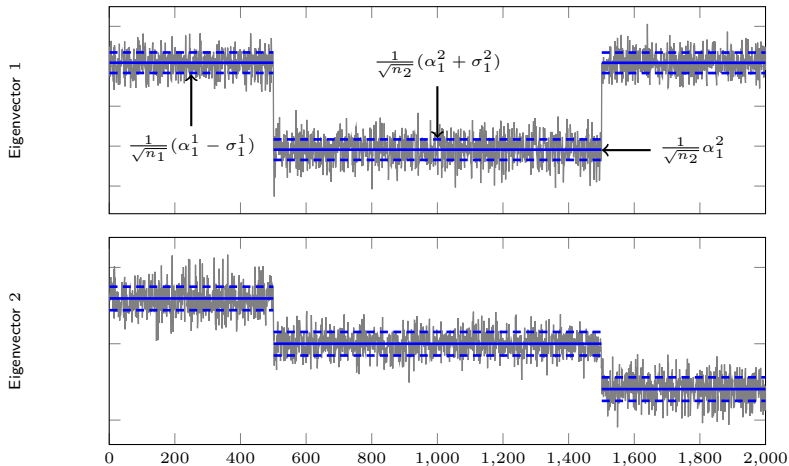
The case  $f'(2) = 0$



**Figure:** Leading two eigenvectors of  $\mathcal{L}$  (or equivalently of  $L$ ) versus deterministic approximations of  $\alpha_i^a \pm \sigma_i^a$ .

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$



**Figure:** Leading two eigenvectors of  $\mathcal{L}$  (or equivalently of  $L$ ) versus deterministic approximations of  $\alpha_i^a \pm \sigma_i^a$ .

# Spectral Clustering: The case $f'(\tau) = 0$

The case  $f'(2) = 0$

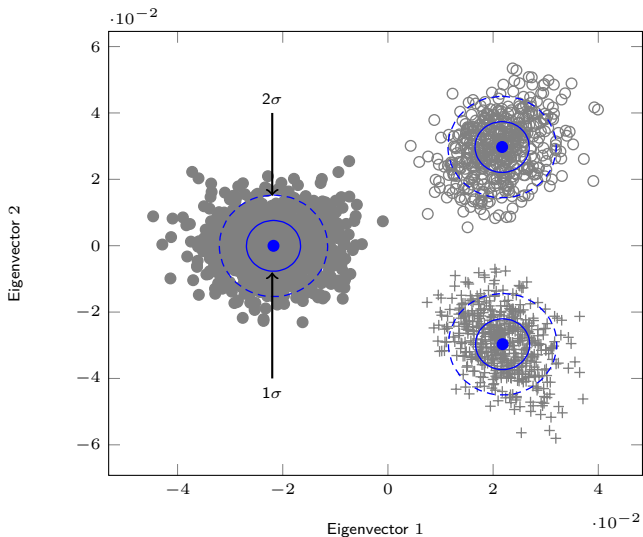


Figure: Leading two eigenvectors of  $\mathcal{L}$  (or equivalently of  $L$ ) versus deterministic approximations of  $\alpha_i^a \pm \sigma_i^a$ .



# Spectral Clustering: The case $f'(\tau) = 0$

Application: Clustering data vectors with close covariances

## Setting.

- ▶  $p$  dimensional vector observations.
- ▶  $m$  data sources.
- ▶  $E[x_i] = 0$ ,  $E[x_i x_i^T] = C_i$ .

# Spectral Clustering: The case $f'(\tau) = 0$

Application: Clustering data vectors with close covariances

## Setting.

- ▶  $p$  dimensional vector observations.
- ▶  $m$  data sources.
- ▶  $E[x_i] = 0$ ,  $E[x_i x_i^T] = C_i$ .
- ▶  $n_i$  independent observations  $x_i^{(1)}, \dots, x_i^{(n_i)}$  for source  $i$ .

# Spectral Clustering: The case $f'(\tau) = 0$

Application: Clustering data vectors with close covariances

## Setting.

- ▶  $p$  dimensional vector observations.
- ▶  $m$  data sources.
- ▶  $E[x_i] = 0$ ,  $E[x_i x_i^T] = C_i$ .
- ▶  $n_i$  independent observations  $x_i^{(1)}, \dots, x_i^{(n_i)}$  for source  $i$ .

**Objective.** Cluster sources based on covariance  $C_i$ .

# Spectral Clustering: The case $f'(\tau) = 0$

Application: Clustering data vectors with close covariances

## Setting.

- ▶  $p$  dimensional vector observations.
- ▶  $m$  data sources.
- ▶  $E[x_i] = 0$ ,  $E[x_i x_i^T] = C_i$ .
- ▶  $n_i$  independent observations  $x_i^{(1)}, \dots, x_i^{(n_i)}$  for source  $i$ .

**Objective.** Cluster sources based on covariance  $C_i$ .

**Applications examples.** Massive MIMO scheduling / EEG classification / etc.

# Spectral Clustering: The case $f'(\tau) = 0$

Application: Clustering data vectors with close covariances

## Setting.

- ▶  $p$  dimensional vector observations.
- ▶  $m$  data sources.
- ▶  $E[x_i] = 0$ ,  $E[x_i x_i^T] = C_i$ .
- ▶  $n_i$  independent observations  $x_i^{(1)}, \dots, x_i^{(n_i)}$  for source  $i$ .

**Objective.** Cluster sources based on covariance  $C_i$ .

**Applications examples.** Massive MIMO scheduling / EEG classification / etc.

## Algorithm.

1. Build kernel matrix  $K$ , then  $\mathcal{L}$ , based on  $mn_i$  vectors  $x_1^{(1)}, \dots, x_m^{(n_i)}$  (as if  $mn_i$  values to cluster).

# Spectral Clustering: The case $f'(\tau) = 0$

Application: Clustering data vectors with close covariances

## Setting.

- ▶  $p$  dimensional vector observations.
- ▶  $m$  data sources.
- ▶  $E[x_i] = 0$ ,  $E[x_i x_i^T] = C_i$ .
- ▶  $n_i$  independent observations  $x_i^{(1)}, \dots, x_i^{(n_i)}$  for source  $i$ .

**Objective.** Cluster sources based on covariance  $C_i$ .

**Applications examples.** Massive MIMO scheduling / EEG classification / etc.

## Algorithm.

1. Build kernel matrix  $K$ , then  $\mathcal{L}$ , based on  $mn_i$  vectors  $x_1^{(1)}, \dots, x_m^{(n_i)}$  (as if  $mn_i$  values to cluster).
2. Extract dominant isolated eigenvectors  $u_1, \dots, u_\kappa$

# Spectral Clustering: The case $f'(\tau) = 0$

Application: Clustering data vectors with close covariances

## Setting.

- ▶  $p$  dimensional vector observations.
- ▶  $m$  data sources.
- ▶  $E[x_i] = 0$ ,  $E[x_i x_i^T] = C_i$ .
- ▶  $n_i$  independent observations  $x_i^{(1)}, \dots, x_i^{(n_i)}$  for source  $i$ .

**Objective.** Cluster sources based on covariance  $C_i$ .

**Applications examples.** Massive MIMO scheduling / EEG classification / etc.

## Algorithm.

1. Build kernel matrix  $K$ , then  $\mathcal{L}$ , based on  $mn_i$  vectors  $x_1^{(1)}, \dots, x_m^{(n_i)}$  (as if  $mn_i$  values to cluster).
2. Extract dominant isolated eigenvectors  $u_1, \dots, u_\kappa$
3. For each  $i$ , create  $\tilde{u}_i = \frac{1}{n_i} (I_m \otimes \mathbf{1}_{n_i}^T) u_i$ , i.e., average eigenvectors along time.

# Spectral Clustering: The case $f'(\tau) = 0$

Application: Clustering data vectors with close covariances

## Setting.

- ▶  $p$  dimensional vector observations.
- ▶  $m$  data sources.
- ▶  $E[x_i] = 0$ ,  $E[x_i x_i^T] = C_i$ .
- ▶  $n_i$  independent observations  $x_i^{(1)}, \dots, x_i^{(n_i)}$  for source  $i$ .

**Objective.** Cluster sources based on covariance  $C_i$ .

**Applications examples.** Massive MIMO scheduling / EEG classification / etc.

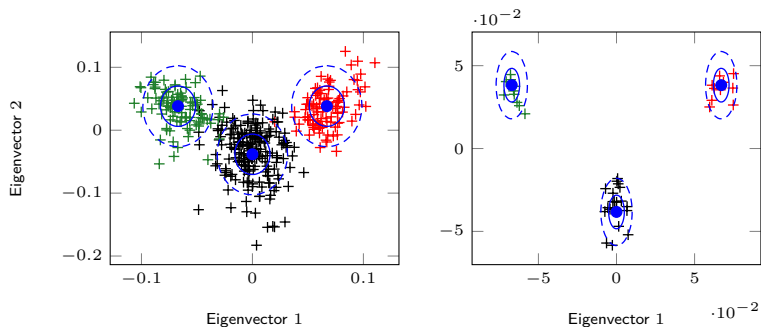
## Algorithm.

1. Build kernel matrix  $K$ , then  $\mathcal{L}$ , based on  $mn_i$  vectors  $x_1^{(1)}, \dots, x_m^{(n_i)}$  (as if  $mn_i$  values to cluster).
2. Extract dominant isolated eigenvectors  $u_1, \dots, u_\kappa$
3. For each  $i$ , create  $\tilde{u}_i = \frac{1}{n_i} (I_m \otimes \mathbf{1}_{n_i}^T) u_i$ , i.e., average eigenvectors along time.
4. Perform  $k$ -class clustering on vectors  $\tilde{u}_1, \dots, \tilde{u}_\kappa$ .



# Spectral Clustering: The case $f'(\tau) = 0$

Application Example: Clustering data vectors with close covariances



**Figure: Clustering data vectors with close covariances application:** Leading two eigenvectors before (left figure) and after (right figure)  $n_i$ -averaging. Setting:  $p = 400$ ,  $m = 40$ ,  $n_i = 10$ ,  $k = 3$ ,  $c_1 = c_3 = 1/4$ ,  $c_2 = 1/2$ . Kernel function  $f(t) = \exp(-(t - 2)^2)$ .

# Spectral Clustering: The case $f'(\tau) = 0$

Application Example: Clustering data vectors with close covariances

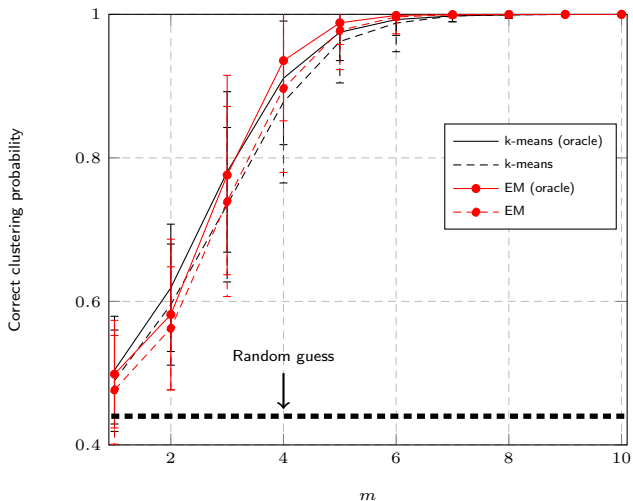


Figure: Overlap for different  $m$ , using the k-means or EM starting from actual centroid solutions (oracle) or randomly.

# Spectral Clustering: The case $f'(\tau) = 0$

Application Example: Clustering data vectors with close covariances

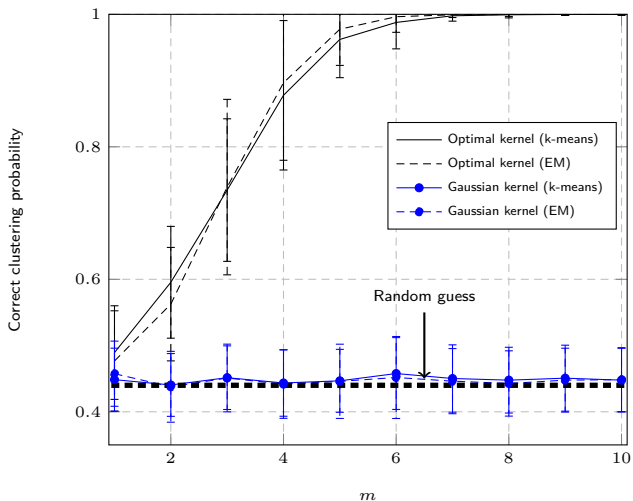


Figure: Overlap for optimal kernel  $f(t)$  (here  $f(t) = \exp(-(t - 2)^2)$ ) and Gaussian kernel  $f(t) = \exp(-t^2)$ , for different  $m$ , using the k-means or EM.

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Optimal growth rates and optimal kernels

## Conclusion of previous analyses:

- ▶ kernel  $f(\frac{1}{p}\|x_i - x_j\|^2)$  with  $f'(\tau) \neq 0$ :
    - ▶ optimal in  $\|\mu_a^\circ\| = O(1)$ ,  $\frac{1}{p}\text{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
    - ▶ suboptimal in  $\frac{1}{p}\text{tr} C_a^\circ C_b^\circ = O(1)$
- **Model type:** Marčenko–Pastur + spikes.

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Optimal growth rates and optimal kernels

## Conclusion of previous analyses:

- ▶ kernel  $f(\frac{1}{p}\|x_i - x_j\|^2)$  with  $f'(\tau) \neq 0$ :
  - ▶ optimal in  $\|\mu_a^\circ\| = O(1)$ ,  $\frac{1}{p}\text{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - ▶ suboptimal in  $\frac{1}{p}\text{tr} C_a^\circ C_b^\circ = O(1)$→ **Model type:** Marčenko–Pastur + spikes.
  
- ▶ kernel  $f(\frac{1}{p}\|x_i - x_j\|^2)$  with  $f'(\tau) = 0$ :
  - ▶ suboptimal in  $\|\mu_a^\circ\| \gg O(1)$  (kills the means)
  - ▶ better in discriminating covariance (stress on  $t$  and  $T$ )→ **Model type:** smaller order semi-circle law + spikes.

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Optimal growth rates and optimal kernels

## Conclusion of previous analyses:

- ▶ kernel  $f(\frac{1}{p}\|x_i - x_j\|^2)$  with  $f'(\tau) \neq 0$ :
  - ▶ optimal in  $\|\mu_a^\circ\| = O(1)$ ,  $\frac{1}{p}\text{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - ▶ suboptimal in  $\frac{1}{p}\text{tr} C_a^\circ C_b^\circ = O(1)$→ **Model type:** Marčenko–Pastur + spikes.
  
- ▶ kernel  $f(\frac{1}{p}\|x_i - x_j\|^2)$  with  $f'(\tau) = 0$ :
  - ▶ suboptimal in  $\|\mu_a^\circ\| \gg O(1)$  (**kills the means**)
  - ▶ better in discriminating covariance (**stress on  $t$  and  $T$** )→ **Model type:** smaller order semi-circle law + spikes.

## Jointly optimal solution:

- ▶ evenly weighing Marčenko–Pastur and semi-circle laws

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Optimal growth rates and optimal kernels

## Conclusion of previous analyses:

- ▶ kernel  $f(\frac{1}{p}\|x_i - x_j\|^2)$  with  $f'(\tau) \neq 0$ :
  - ▶ optimal in  $\|\mu_a^\circ\| = O(1)$ ,  $\frac{1}{p}\text{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - ▶ suboptimal in  $\frac{1}{p}\text{tr} C_a^\circ C_b^\circ = O(1)$→ **Model type:** Marčenko–Pastur + spikes.
  
- ▶ kernel  $f(\frac{1}{p}\|x_i - x_j\|^2)$  with  $f'(\tau) = 0$ :
  - ▶ suboptimal in  $\|\mu_a^\circ\| \gg O(1)$  (kills the means)
  - ▶ better in discriminating covariance (stress on  $t$  and  $T$ )→ **Model type:** smaller order semi-circle law + spikes.

## Jointly optimal solution:

- ▶ evenly weighing Marčenko–Pastur and semi-circle laws
- ▶ the “ $\alpha$ - $\beta$ ” kernel:

$$f'(\tau) = \frac{\alpha}{\sqrt{p}}, \quad \frac{1}{2}f''(\tau) = \beta.$$

## Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

New assumption setting

- ▶ We consider now an **improved growth rate setting**.

### Assumption (Optimal Growth Rate)

As  $n \rightarrow \infty$ ,

1. **Data scaling:**  $\frac{p}{n} \rightarrow c_0 \in (0, \infty)$ ,  $\frac{n_a}{n} \rightarrow c_a \in (0, 1)$ ,
2. **Mean scaling:** with  $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$  and  $\mu_a^\circ \triangleq \mu_a - \mu^\circ$ , then  $\|\mu_a^\circ\| = O(1)$
3. **Covariance scaling:** with  $C^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} C_a$  and  $C_a^\circ \triangleq C_a - C^\circ$ , then

$$\|C_a\| = O(1), \quad \text{tr} C_a^\circ = O(\sqrt{p}), \quad \text{tr} C_a^\circ C_b^\circ = O(\sqrt{p}).$$



# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

New assumption setting

- ▶ We consider now an **improved growth rate setting**.

## Assumption (Optimal Growth Rate)

As  $n \rightarrow \infty$ ,

1. **Data scaling:**  $\frac{p}{n} \rightarrow c_0 \in (0, \infty)$ ,  $\frac{n_a}{n} \rightarrow c_a \in (0, 1)$ ,
2. **Mean scaling:** with  $\mu^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} \mu_a$  and  $\mu_a^\circ \triangleq \mu_a - \mu^\circ$ , then  $\|\mu_a^\circ\| = O(1)$
3. **Covariance scaling:** with  $C^\circ \triangleq \sum_{a=1}^k \frac{n_a}{n} C_a$  and  $C_a^\circ \triangleq C_a - C^\circ$ , then

$$\|C_a\| = O(1), \quad \text{tr} C_a^\circ = O(\sqrt{p}), \quad \text{tr} C_a^\circ C_b^\circ = O(\sqrt{p}).$$

**Kernel:**

- ▶ For technical simplicity, we consider

$$\tilde{K} = PKP = P \left\{ f \left( \frac{1}{p} (x^{(i)})^\top (x_j^{(i)}) \right) \right\}_{i,j=1}^n P, \quad P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

i.e.,  $\tau$  replaced by 0.

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

## Main Results

### Theorem

As  $n \rightarrow \infty$ ,

$$\left\| \sqrt{p} \left( PKP + (f(0) + \tau f'(0)) P \right) - \hat{\mathcal{K}} \right\| \xrightarrow{\text{a.s.}} 0$$

with, for  $\alpha = \sqrt{p}f'(0) = O(1)$  and  $\beta = \frac{1}{2}f''(0) = O(1)$ ,

$$\hat{\mathcal{K}} = \alpha PW^T WP + \beta P\Phi P + UAU^T$$

$$A = \begin{bmatrix} \alpha M^T M + \beta T & \alpha I_k \\ \alpha I_k & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} J \\ \frac{1}{\sqrt{p}}, PW^T M \end{bmatrix}$$

$$\frac{\Phi}{\sqrt{p}} = \left\{ ((\omega_i^\circ)^\top \omega_j^\circ)^2 \delta_{i \neq j} \right\}_{i,j=1}^n - \left\{ \frac{\text{tr}(C_a C_b)}{p^2} \mathbf{1}_{n_a} \mathbf{1}_{n_b}^\top \right\}_{a,b=1}^k.$$

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

## Main Results

### Theorem

As  $n \rightarrow \infty$ ,

$$\left\| \sqrt{p} \left( PKP + (f(0) + \tau f'(0)) P \right) - \hat{\mathcal{K}} \right\| \xrightarrow{\text{a.s.}} 0$$

with, for  $\alpha = \sqrt{p}f'(0) = O(1)$  and  $\beta = \frac{1}{2}f''(0) = O(1)$ ,

$$\hat{\mathcal{K}} = \alpha PW^T WP + \beta P\Phi P + UAU^T$$

$$A = \begin{bmatrix} \alpha M^T M + \beta T & \alpha I_k \\ \alpha I_k & 0 \end{bmatrix}$$

$$U = \begin{bmatrix} J \\ \frac{1}{\sqrt{p}}, PW^T M \end{bmatrix}$$

$$\frac{\Phi}{\sqrt{p}} = \left\{ ((\omega_i^\circ)^\top \omega_j^\circ)^2 \delta_{i \neq j} \right\}_{i,j=1}^n - \left\{ \frac{\text{tr}(C_a C_b)}{p^2} 1_{n_a} 1_{n_b}^\top \right\}_{a,b=1}^k.$$

**Role of  $\alpha, \beta$ :**

- ▶ Weighs **Marčenko–Pastur** versus **semi-circle** parts.

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Limiting eigenvalue distribution

## Theorem (Eigenvalues Bulk)

As  $p \rightarrow \infty$ ,

$$\nu_n \triangleq \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(\hat{K})} \xrightarrow{\text{a.s.}} \nu$$

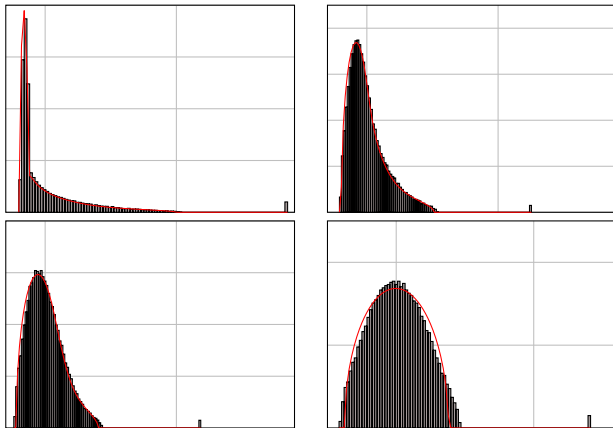
with  $\nu$  having Stieltjes transform  $m(z)$  solution of

$$\frac{1}{m(z)} = -z + \frac{\alpha}{p} \text{tr} C^\circ \left( I_k + \frac{\alpha m(z)}{c_0} C^\circ \right)^{-1} - \frac{2\beta^2}{c_0} \omega^2 m(z)$$

where  $\omega = \lim_{p \rightarrow \infty} \frac{1}{p} \text{tr}(C^\circ)^2$ .

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Limiting eigenvalue distribution



**Figure:** Eigenvalues of  $K$  (up to recentering) versus limiting law,  $p = 2048$ ,  $n = 4096$ ,  $k = 2$ ,  $n_1 = n_2$ ,  $\mu_i = 3\delta_i$ ,  $f(x) = \frac{1}{2}\beta \left(x + \frac{1}{\sqrt{p}} \frac{\alpha}{\beta}\right)^2$ . **(Top left):**  $\alpha = 8, \beta = 1$ , **(Top right):**  $\alpha = 4, \beta = 3$ , **(Bottom left):**  $\alpha = 3, \beta = 4$ , **(Bottom right):**  $\alpha = 1, \beta = 8$ .

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Asymptotic performances: MNIST

- ▶ MNIST is “means-dominant” but not that much!

DATASETS	$\ \mu_1^\circ - \mu_2^\circ\ ^2$	$\frac{1}{\sqrt{p}} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$
MNIST (DIGITS 1, 7)	613	1990	71.1
MNIST (DIGITS 3, 6)	441	1119	39.9
MNIST (DIGITS 3, 8)	212	652	23.5

# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Asymptotic performances: MNIST

- ▶ MNIST is “means-dominant” but not that much!

DATASETS	$\ \mu_1^\circ - \mu_2^\circ\ ^2$	$\frac{1}{\sqrt{p}} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$
MNIST (DIGITS 1, 7)	613	1990	71.1
MNIST (DIGITS 3, 6)	441	1119	39.9
MNIST (DIGITS 3, 8)	212	652	23.5

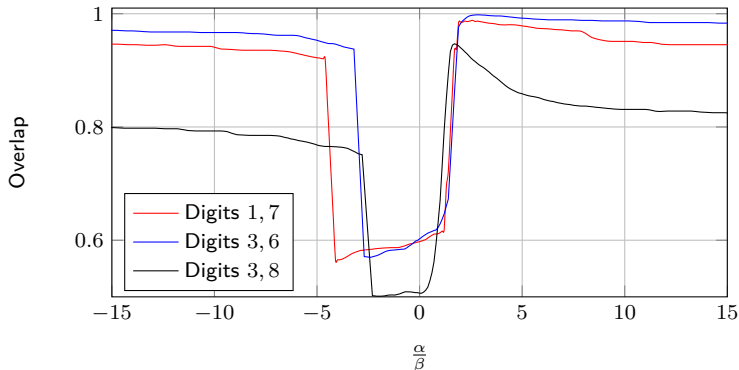


Figure: Spectral clustering of the MNIST database for varying  $\frac{\alpha}{\beta}$ .

## Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Asymptotic performances: EEG data

- ▶ EEG data are “variance-dominant”

DATASETS	$\ \boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\ ^2$	$\frac{1}{\sqrt{p}} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$
EEG (SETS $A, E$ )	2.4	10.9	1.1



# Kernel Spectral Clustering: The case $f'(\tau) = \frac{\alpha}{\sqrt{p}}$

Asymptotic performances: EEG data

- ▶ EEG data are “variance-dominant”

DATASETS	$\ \mu_1^\circ - \mu_2^\circ\ ^2$	$\frac{1}{\sqrt{p}} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \text{TR}(\mathbf{C}_1 - \mathbf{C}_2)^2$
EEG (SETS $A, E$ )	2.4	10.9	1.1

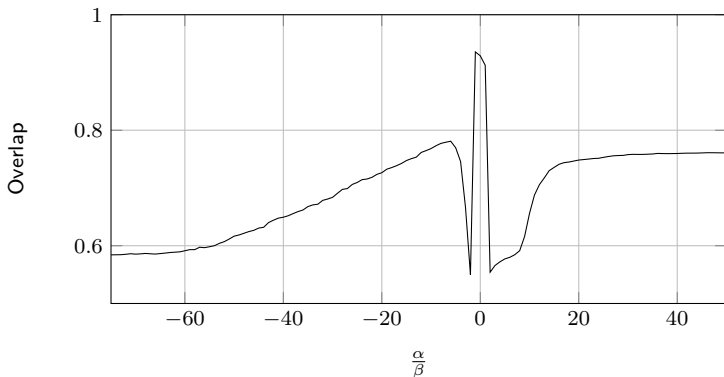


Figure: Spectral clustering of the EEG database for varying  $\frac{\alpha}{\beta}$ .

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

Semi-Supervised Learning

From Gaussian Mixtures to Real Data

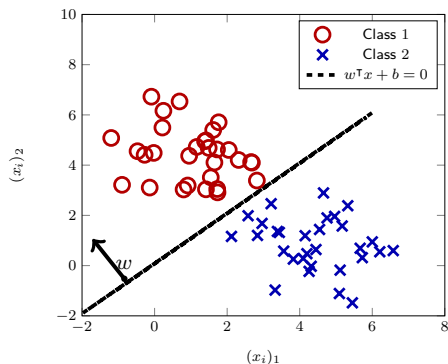
## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

# LS-SVM Problem Statement

Optimization problem: find separating hyperplane (linear separability case)

$$\arg \min_w J(w, e) = \|w\|^2 + \frac{\gamma}{n} \sum_{i=1}^n e_i^2$$

$$\text{such that } y_i = w^T x_i + b + e_i \\ \text{for } i = 1, \dots, n$$

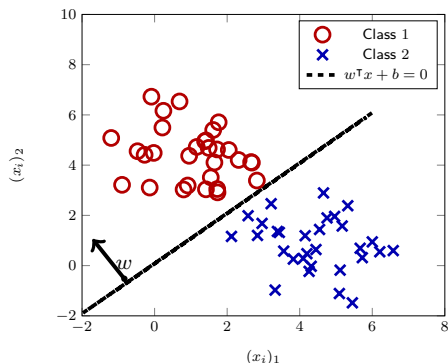


# LS-SVM Problem Statement

Optimization problem: find separating hyperplane (linear separability case)

$$\arg \min_w J(w, e) = \|w\|^2 + \frac{\gamma}{n} \sum_{i=1}^n e_i^2$$

$$\text{such that } y_i = w^T x_i + b + e_i \\ \text{for } i = 1, \dots, n$$



## Advantage of LS-SVM

Explicit form, as opposed to SVM  $\Rightarrow$  easier to analyze.

## LS-SVM Problem Statement

When **no linear separability**:

⇒ Kernel method

# LS-SVM Problem Statement

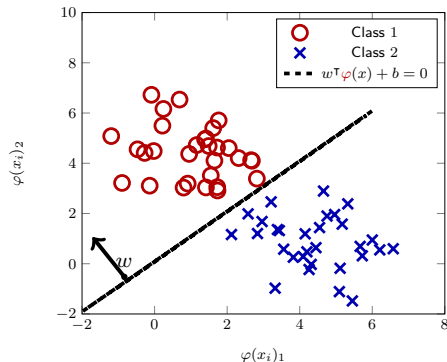
When **no linear separability**:

⇒ Kernel method

To solve the optimization problem:

$$\arg \min_w J(w, e) = \|w\|^2 + \frac{\gamma}{n} \sum_{i=1}^n e_i^2$$

$$\text{such that } y_i = w^\top \varphi(x_i) + b + e_i \\ \text{for } i = 1, \dots, n$$



- **Training:** Solution given by  $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ , where

$$\begin{cases} \alpha &= S \left( I_n - \frac{1_n 1_n^\top S}{1_n^\top S 1_n} \right) y = S (y - b 1_n) \\ b &= \frac{1_n^\top S y}{1_n^\top S 1_n} \end{cases}$$

with  $S \equiv \left( K + \frac{n}{\gamma} I_n \right)^{-1}$  resolvent of **kernel matrix**:

$$K \equiv \{ \varphi(x_i)^\top \varphi(x_j) \}_{i,j=1}^n$$



- **Training:** Solution given by  $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ , where

$$\begin{cases} \alpha &= S \left( I_n - \frac{1_n 1_n^T S}{1_n^T S 1_n} \right) y = S (y - b 1_n) \\ b &= \frac{1_n^T S y}{1_n^T S 1_n} \end{cases}$$

with  $S \equiv \left( K + \frac{n}{\gamma} I_n \right)^{-1}$  resolvent of **kernel matrix**:

$$K \equiv \{ \varphi(x_i)^T \varphi(x_j) \}_{i,j=1}^n$$

- **Training:** Solution given by  $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ , where

$$\begin{cases} \alpha &= S \left( I_n - \frac{1_n 1_n^\top S}{1_n^\top S 1_n} \right) y = S (y - b 1_n) \\ b &= \frac{1_n^\top S y}{1_n^\top S 1_n} \end{cases}$$

with  $S \equiv \left( K + \frac{n}{\gamma} I_n \right)^{-1}$  resolvent of **kernel matrix**:

$$K \equiv \{\varphi(x_i)^\top \varphi(x_j)\}_{i,j=1}^n \underbrace{=}_{\text{kernel trick}} \left\{ f \left( \frac{\|x_i - x_j\|^2}{p} \right) \right\}_{i,j=1}^n$$

for some *translation invariant kernel function*  $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ ,  $y \equiv [y_1, \dots, y_n]^\top$  and  $\alpha \equiv [\alpha_1, \dots, \alpha_n]^\top$ .

- ▶ **Training:** Solution given by  $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ , where

$$\begin{cases} \alpha &= S \left( I_n - \frac{1_n 1_n^\top S}{1_n^\top S 1_n} \right) y = S (y - b 1_n) \\ b &= \frac{1_n^\top S y}{1_n^\top S 1_n} \end{cases}$$

with  $S \equiv \left( K + \frac{n}{\gamma} I_n \right)^{-1}$  resolvent of **kernel matrix**:

$$K \equiv \underbrace{\{\varphi(x_i)^\top \varphi(x_j)\}_{i,j=1}^n}_{\text{kernel trick}} \equiv \left\{ f \left( \frac{\|x_i - x_j\|^2}{p} \right) \right\}_{i,j=1}^n$$

for some *translation invariant kernel function*  $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ ,  $y \equiv [y_1, \dots, y_n]^\top$  and  $\alpha \equiv [\alpha_1, \dots, \alpha_n]^\top$ .

- ▶ **Inference:** **Decision** for new  $x$

$$g(x) = \alpha^\top k(x) + b \text{ where } k(x) = \left\{ f \left( \|x_j - x\|^2/p \right) \right\}_{j=1}^n \in \mathbb{R}^n$$

- ▶ **Training:** Solution given by  $w = \sum_{i=1}^n \alpha_i \varphi(x_i)$ , where

$$\begin{cases} \alpha &= S \left( I_n - \frac{1_n 1_n^\top S}{1_n^\top S 1_n} \right) y = S (y - b 1_n) \\ b &= \frac{1_n^\top S y}{1_n^\top S 1_n} \end{cases}$$

with  $S \equiv \left( K + \frac{n}{\gamma} I_n \right)^{-1}$  resolvent of **kernel matrix**:

$$K \equiv \underbrace{\{\varphi(x_i)^\top \varphi(x_j)\}_{i,j=1}^n}_{\text{kernel trick}} \equiv \left\{ f \left( \frac{\|x_i - x_j\|^2}{p} \right) \right\}_{i,j=1}^n$$

for some *translation invariant kernel function*  $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ ,  $y \equiv [y_1, \dots, y_n]^\top$  and  $\alpha \equiv [\alpha_1, \dots, \alpha_n]^\top$ .

- ▶ **Inference:** **Decision** for new  $x$

$$g(x) = \alpha^\top k(x) + b \text{ where } k(x) = \left\{ f \left( \|x_j - x\|^2/p \right) \right\}_{j=1}^n \in \mathbb{R}^n$$

- ▶ In practice, **sign**( $g(x)$ ) to predict the class.

- ▶ **Large dimension:**  $n, p \rightarrow \infty$  and  $\frac{p}{n} \rightarrow c_0$

- ▶ **Large dimension:**  $n, p \rightarrow \infty$  and  $\frac{p}{n} \rightarrow c_0$
- ▶ **Gaussian mixture model:** for  $a \in \{1, 2\}$ :

$$x_i \sim \mathcal{N}(\mu_a, C_a)$$

- ▶ **Large dimension:**  $n, p \rightarrow \infty$  and  $\frac{p}{n} \rightarrow c_0$
- ▶ **Gaussian mixture model:** for  $a \in \{1, 2\}$ :

$$x_i \sim \mathcal{N}(\mu_a, C_a)$$

- ▶ **Non-trivial regime:** to ensure  $P(x_i \rightarrow \mathcal{C}_b \mid x_i \in \mathcal{C}_a) \not\rightarrow 0$  nor  $1$

- ▶ **Large dimension:**  $n, p \rightarrow \infty$  and  $\frac{p}{n} \rightarrow c_0$
- ▶ **Gaussian mixture model:** for  $a \in \{1, 2\}$ :

$$x_i \sim \mathcal{N}(\mu_a, C_a)$$

- ▶ **Non-trivial regime:** to ensure  $P(x_i \rightarrow \mathcal{C}_b \mid x_i \in \mathcal{C}_a) \not\rightarrow 0$  nor 1
  - ▶  $\|\mu_2 - \mu_1\| = \mathcal{O}(1)$



- ▶ **Large dimension:**  $n, p \rightarrow \infty$  and  $\frac{p}{n} \rightarrow c_0$
- ▶ **Gaussian mixture model:** for  $a \in \{1, 2\}$ :

$$x_i \sim \mathcal{N}(\mu_a, C_a)$$

- ▶ **Non-trivial regime:** to ensure  $P(x_i \rightarrow \mathcal{C}_b \mid x_i \in \mathcal{C}_a) \not\rightarrow 0$  nor 1
  - ▶  $\|\mu_2 - \mu_1\| = \mathcal{O}(1)$
  - ▶  $\|C_a\| = \mathcal{O}(1)$  and  $\text{tr}(C_2 - C_1) = \mathcal{O}(\sqrt{n})$

- ▶ **Large dimension:**  $n, p \rightarrow \infty$  and  $\frac{p}{n} \rightarrow c_0$
- ▶ **Gaussian mixture model:** for  $a \in \{1, 2\}$ :

$$x_i \sim \mathcal{N}(\mu_a, C_a)$$

- ▶ **Non-trivial regime:** to ensure  $P(x_i \rightarrow \mathcal{C}_b \mid x_i \in \mathcal{C}_a) \not\rightarrow 0$  nor 1
  - ▶  $\|\mu_2 - \mu_1\| = \mathcal{O}(1)$
  - ▶  $\|C_a\| = \mathcal{O}(1)$  and  $\text{tr}(C_2 - C_1) = \mathcal{O}(\sqrt{n})$
- ▶ **Notations:**

- ▶ **Large dimension:**  $n, p \rightarrow \infty$  and  $\frac{p}{n} \rightarrow c_0$
- ▶ **Gaussian mixture model:** for  $a \in \{1, 2\}$ :

$$x_i \sim \mathcal{N}(\mu_a, C_a)$$

- ▶ **Non-trivial regime:** to ensure  $P(x_i \rightarrow \mathcal{C}_b \mid x_i \in \mathcal{C}_a) \not\rightarrow 0$  nor 1
  - ▶  $\|\mu_2 - \mu_1\| = \mathcal{O}(1)$
  - ▶  $\|C_a\| = \mathcal{O}(1)$  and  $\text{tr}(C_2 - C_1) = \mathcal{O}(\sqrt{n})$
- ▶ **Notations:**
  - ▶  $C^\circ \equiv c_1 C_1 + c_2 C_2$ ,  $c_1 \equiv \frac{n_1}{n}$  and  $c_2 \equiv \frac{n_2}{n} = 1 - c_1$

- ▶ **Large dimension:**  $n, p \rightarrow \infty$  and  $\frac{p}{n} \rightarrow c_0$
- ▶ **Gaussian mixture model:** for  $a \in \{1, 2\}$ :

$$x_i \sim \mathcal{N}(\mu_a, C_a)$$

- ▶ **Non-trivial regime:** to ensure  $P(x_i \rightarrow \mathcal{C}_b \mid x_i \in \mathcal{C}_a) \not\rightarrow 0$  nor 1
  - ▶  $\|\mu_2 - \mu_1\| = \mathcal{O}(1)$
  - ▶  $\|C_a\| = \mathcal{O}(1)$  and  $\text{tr}(C_2 - C_1) = \mathcal{O}(\sqrt{n})$
- ▶ **Notations:**
  - ▶  $C^\circ \equiv c_1 C_1 + c_2 C_2$ ,  $c_1 \equiv \frac{n_1}{n}$  and  $c_2 \equiv \frac{n_2}{n} = 1 - c_1$
  - ▶ **Key Notation:**  $\tau \equiv \frac{2}{p} \text{tr} C^\circ$

## Reminder: kernel matrix

$$K_{i,j} = f\left(\frac{\|x_i - x_j\|^2}{p}\right)$$

For  $x_i \in \mathcal{C}_a$  and  $x_j \in \mathcal{C}_b$ :  $\frac{1}{p}\|x_i - x_j\|^2 = \tau + \mathcal{O}(n^{-1/2})$ , thus for  $K_{i,j}$

$$K_{i,j} = f\left(\tau + \mathcal{O}(n^{-1/2})\right) = f(\tau) + f'(\tau)[\dots] + f''(\tau)[\dots] + \dots$$

or in matrix form

$$K = f(\tau)1_n 1_n^T + f'(\tau)[\dots] + f''(\tau)[\dots] + \dots$$

## Reminder: kernel matrix

$$K_{i,j} = f\left(\frac{\|x_i - x_j\|^2}{p}\right)$$

For  $x_i \in \mathcal{C}_a$  and  $x_j \in \mathcal{C}_b$ :  $\frac{1}{p}\|x_i - x_j\|^2 = \tau + \mathcal{O}(n^{-1/2})$ , thus for  $K_{i,j}$

$$K_{i,j} = f\left(\tau + \mathcal{O}(n^{-1/2})\right) = f(\tau) + f'(\tau)[\dots] + f''(\tau)[\dots] + \dots$$

or in matrix form

$$K = f(\tau)1_n 1_n^\top + f'(\tau)[\dots] + f''(\tau)[\dots] + \dots$$

## Consequence

Asymptotic statistics of  $K$ , **thus of**

$$g(x) = \alpha^\top k(x) + b$$

## Reminder: kernel matrix

$$K_{i,j} = f\left(\frac{\|x_i - x_j\|^2}{p}\right)$$

For  $x_i \in \mathcal{C}_a$  and  $x_j \in \mathcal{C}_b$ :  $\frac{1}{p}\|x_i - x_j\|^2 = \tau + \mathcal{O}(n^{-1/2})$ , thus for  $K_{i,j}$

$$K_{i,j} = f\left(\tau + \mathcal{O}(n^{-1/2})\right) = f(\tau) + f'(\tau)[\dots] + f''(\tau)[\dots] + \dots$$

or in matrix form

$$K = f(\tau)1_n 1_n^\top + f'(\tau)[\dots] + f''(\tau)[\dots] + \dots$$

## Consequence

Asymptotic statistics of  $K$ , **thus of**

$$g(x) = \alpha^\top k(x) + b$$

$$\begin{cases} \alpha &= S \left( I_n - \frac{1_n 1_n^\top S}{1_n^\top S 1_n} \right) y = S (y - b 1_n) \\ b &= \frac{1_n^\top S y}{1_n^\top S 1_n} \end{cases}, \quad S \equiv \left( K + \frac{n}{\gamma} I_n \right)^{-1}$$

## Asymptotic Behavior of the Decision Function

### Theorem ([Liao,C'19])

Under previous assumptions, for  $x \in \mathcal{C}_a$ ,  $a \in \{1, 2\}$

$$n(g(x) - G_a) \xrightarrow{d} 0$$

where  $G_a \sim \mathcal{N}(E_a, \text{Var}_a)$



# Asymptotic Behavior of the Decision Function

## Theorem ([Liao,C'19])

Under previous assumptions, for  $x \in \mathcal{C}_a$ ,  $a \in \{1, 2\}$

$$n(g(x) - G_a) \xrightarrow{d} 0$$

where  $G_a \sim \mathcal{N}(E_a, \text{Var}_a)$  with

$$E_a = \begin{cases} c_2 - c_1 - \frac{2}{p} c_2 \cdot c_1 c_2 \gamma \mathfrak{D}, & a = 1 \\ c_2 - c_1 + \frac{2}{p} 2c_1 \cdot c_1 c_2 \gamma \mathfrak{D}, & a = 2 \end{cases}$$
$$\text{Var}_a = \frac{8}{p^2} \gamma^2 c_1^2 c_2^2 (\mathcal{V}_1^a + \mathcal{V}_2^a + \mathcal{V}_3^a)$$

# Asymptotic Behavior of the Decision Function

## Theorem ([Liao,C'19])

Under previous assumptions, for  $x \in \mathcal{C}_a$ ,  $a \in \{1, 2\}$

$$n(g(x) - G_a) \xrightarrow{d} 0$$

where  $G_a \sim \mathcal{N}(E_a, \text{Var}_a)$  with

$$E_a = \begin{cases} c_2 - c_1 - \frac{2}{p} c_2 \cdot c_1 c_2 \gamma \mathfrak{D}, & a = 1 \\ c_2 - c_1 + \frac{2}{p} 2c_1 \cdot c_1 c_2 \gamma \mathfrak{D}, & a = 2 \end{cases}$$
$$\text{Var}_a = \frac{8}{p^2} \gamma^2 c_1^2 c_2^2 (\mathcal{V}_1^a + \mathcal{V}_2^a + \mathcal{V}_3^a)$$

and

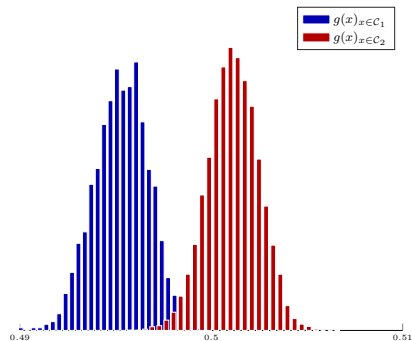
$$\mathfrak{D} = -2f'(\tau) \|\mu_2 - \mu_1\|^2 + \frac{f''(\tau)}{p} (\text{tr}(C_2 - C_1))^2 + \frac{2f''(\tau)}{p} \text{tr}((C_2 - C_1)^2)$$

$$\mathcal{V}_1^a = \frac{(f''(\tau))^2}{p^2} (\text{tr}(C_2 - C_1))^2 \text{tr} C_a^2$$

$$\mathcal{V}_2^a = 2(f'(\tau))^2 (\mu_2 - \mu_1)^\top C_a (\mu_2 - \mu_1)$$

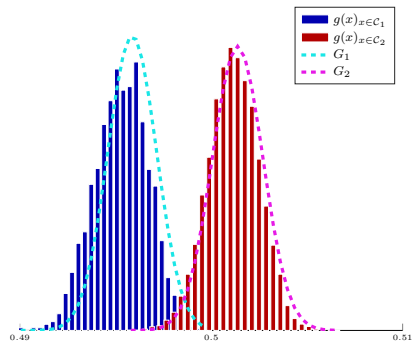
$$\mathcal{V}_3^a = \frac{2(f'(\tau))^2}{n} \left( \frac{\text{tr} C_1 C_a}{c_1} + \frac{\text{tr} C_2 C_a}{c_2} \right)$$

## Simulations on Gaussian data



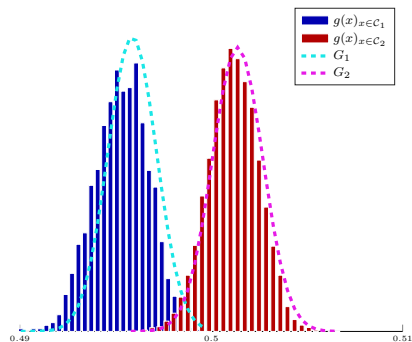
**Figure:** Gaussian approximation of  $g(x)$ ,  
 $n = 256$ ,  $p = 512$ ,  $c_1 = 1/4$ ,  $c_2 = 3/4$ ,  $\gamma = 1$ ,  
Gaussian kernel with  $\sigma^2 = 1$ ,  $x \sim \mathcal{N}(\mu_a, C_a)$   
with  $\mu_a = [0_{a-1}; 3; 0_{p-a}]$ ,  $C_1 = I_p$  and  
 $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{5}{\sqrt{p}})$ .

## Simulations on Gaussian data

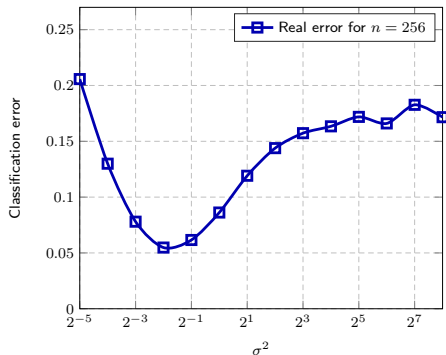


**Figure:** Gaussian approximation of  $g(x)$ ,  
 $n = 256$ ,  $p = 512$ ,  $c_1 = 1/4$ ,  $c_2 = 3/4$ ,  $\gamma = 1$ ,  
Gaussian kernel with  $\sigma^2 = 1$ ,  $x \sim \mathcal{N}(\mu_a, C_a)$   
with  $\mu_a = [0_{a-1}; 3; 0_{p-a}]$ ,  $C_1 = I_p$  and  
 $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{5}{\sqrt{p}})$ .

## Simulations on Gaussian data

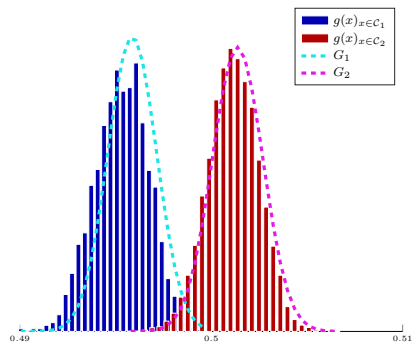


**Figure:** Gaussian approximation of  $g(x)$ ,  $n = 256$ ,  $p = 512$ ,  $c_1 = 1/4$ ,  $c_2 = 3/4$ ,  $\gamma = 1$ , Gaussian kernel with  $\sigma^2 = 1$ ,  $x \sim \mathcal{N}(\mu_a, C_a)$  with  $\mu_a = [0_{a-1}; 3; 0_{p-a}]$ ,  $C_1 = I_p$  and  $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{5}{\sqrt{p}})$ .

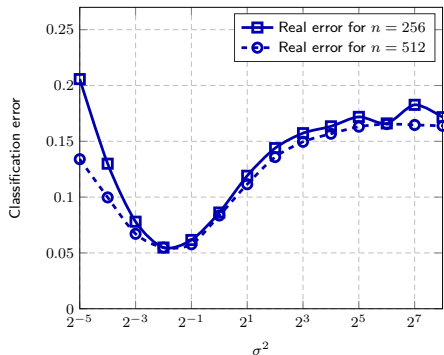


**Figure:** Performance of LS-SVM,  $c_0 = 2$ ,  $c_1 = c_2 = 1/2$ ,  $\gamma = 1$ , Gaussian kernel  $f(t) = \exp(-\frac{t}{2\sigma^2})$ .  $x \sim \mathcal{N}(\mu_a, C_a)$ , with  $\mu_a = [0_{a-1}; 2; 0_{p-a}]$ ,  $C_1 = I_p$  and  $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{4}{\sqrt{p}})$ .

## Simulations on Gaussian data

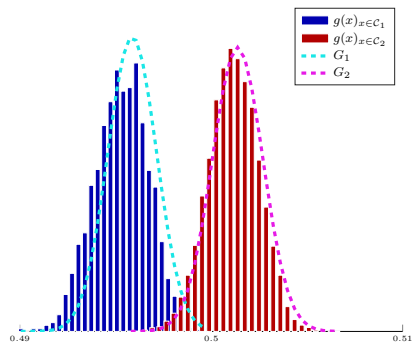


**Figure:** Gaussian approximation of  $g(x)$ ,  $n = 256$ ,  $p = 512$ ,  $c_1 = 1/4$ ,  $c_2 = 3/4$ ,  $\gamma = 1$ , Gaussian kernel with  $\sigma^2 = 1$ ,  $x \sim \mathcal{N}(\mu_a, C_a)$  with  $\mu_a = [0_{a-1}; 3; 0_{p-a}]$ ,  $C_1 = I_p$  and  $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{5}{\sqrt{p}})$ .

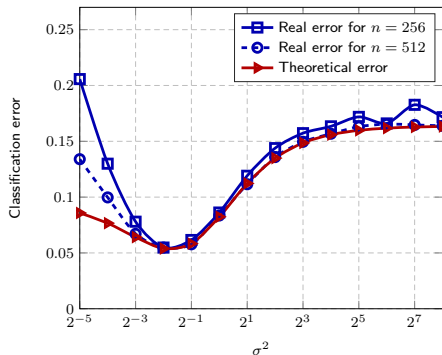


**Figure:** Performance of LS-SVM,  $c_0 = 2$ ,  $c_1 = c_2 = 1/2$ ,  $\gamma = 1$ , Gaussian kernel  $f(t) = \exp(-\frac{t}{2\sigma^2})$ .  $x \sim \mathcal{N}(\mu_a, C_a)$ , with  $\mu_a = [0_{a-1}; 2; 0_{p-a}]$ ,  $C_1 = I_p$  and  $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{4}{\sqrt{p}})$ .

## Simulations on Gaussian data

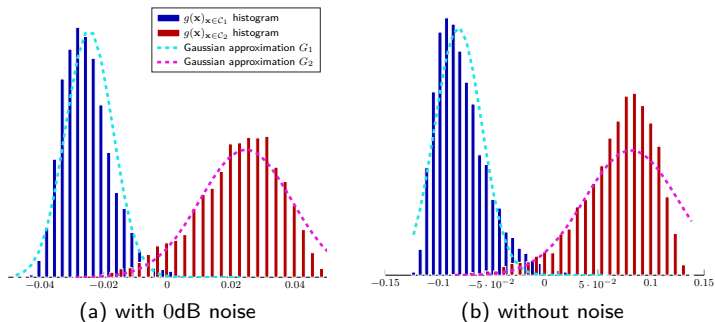


**Figure:** Gaussian approximation of  $g(x)$ ,  $n = 256$ ,  $p = 512$ ,  $c_1 = 1/4$ ,  $c_2 = 3/4$ ,  $\gamma = 1$ , Gaussian kernel with  $\sigma^2 = 1$ ,  $x \sim \mathcal{N}(\mu_a, C_a)$  with  $\mu_a = [0_{a-1}; 3; 0_{p-a}]$ ,  $C_1 = I_p$  and  $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{5}{\sqrt{p}})$ .



**Figure:** Performance of LS-SVM,  $c_0 = 2$ ,  $c_1 = c_2 = 1/2$ ,  $\gamma = 1$ , Gaussian kernel  $f(t) = \exp(-\frac{t}{2\sigma^2})$ .  $x \sim \mathcal{N}(\mu_a, C_a)$ , with  $\mu_a = [0_{a-1}; 2; 0_{p-a}]$ ,  $C_1 = I_p$  and  $\{C_2\}_{i,j} = .4^{|i-j|}(1 + \frac{4}{\sqrt{p}})$ .

## Simulations on MNIST data



**Figure:** Gaussian approximation of  $g(\mathbf{x})$ ,  $n = 256$ ,  $p = 784$ ,  $c_1 = c_2 = 1/2$ ,  $\gamma = 1$ , Gaussian kernel with  $\sigma^2 = 1$ , MNIST data (numbers 1 and 7) without and with 0dB noise.



## Basics of Random Matrix Theory (**Romain COUILLET**)

Motivation: Large Sample Covariance Matrices

The Stieltjes Transform Method

Spiked Models

Other Common Random Matrix Models

Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

*Motivation: EEG-based clustering*

Covariance Distance Inference

*Revisiting Motivation*

Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

Support Vector Machines

**Semi-Supervised Learning**

From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## SSL Problem Statement

**Context:** Similar to clustering:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k$  classes, with  $n_l$  **labelled** and  $n_u$  **unlabelled** data.

## SSL Problem Statement

**Context:** Similar to clustering:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k$  classes, with  $n_l$  **labelled** and  $n_u$  **unlabelled** data.
- ▶ Problem statement: **give scores**  $F_{ia}$  ( $d_i = [K1_n]_i$ )

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^k \sum_{i,j} K_{ij} (F_{ia} - F_{ja})^2$$

such that  $F_{ia} = \delta_{\{x_i \in C_a\}}$ , for all labelled  $x_i$ .

## SSL Problem Statement

**Context:** Similar to clustering:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k$  classes, with  $n_l$  **labelled** and  $n_u$  **unlabelled** data.
- ▶ Problem statement: **give scores**  $F_{ia}$  ( $d_i = [K1_n]_i$ )

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^k \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

such that  $F_{ia} = \delta_{\{x_i \in C_a\}}$ , for all labelled  $x_i$ .

## SSL Problem Statement

**Context:** Similar to clustering:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k$  classes, with  $n_l$  **labelled** and  $n_u$  **unlabelled** data.
- ▶ Problem statement: **give scores**  $F_{ia}$  ( $d_i = [K1_n]_i$ )

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^k \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

such that  $F_{ia} = \delta_{\{x_i \in C_a\}}$ , for all labelled  $x_i$ .

- ▶ **Solution:** for  $F^{(u)} \in \mathbb{R}^{n_u \times k}$ ,  $F^{(l)} \in \mathbb{R}^{n_l \times k}$  scores of unlabelled/labelled data,

$$F^{(u)} = \left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

where we naturally decompose

$$K = \begin{bmatrix} K^{(l,l)} & K^{(l,u)} \\ K^{(u,l)} & K^{(u,u)} \end{bmatrix}$$
$$D = \begin{bmatrix} D^{(l)} & 0 \\ 0 & D^{(u)} \end{bmatrix} = \operatorname{diag} \{K1_n\}.$$

## The finite-dimensional intuition: What we expect

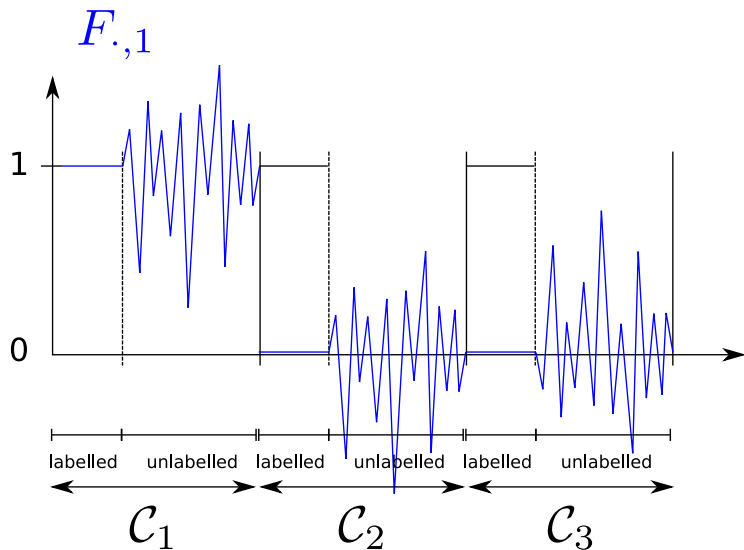


Figure: Typical expected performance output

## The finite-dimensional intuition: What we expect

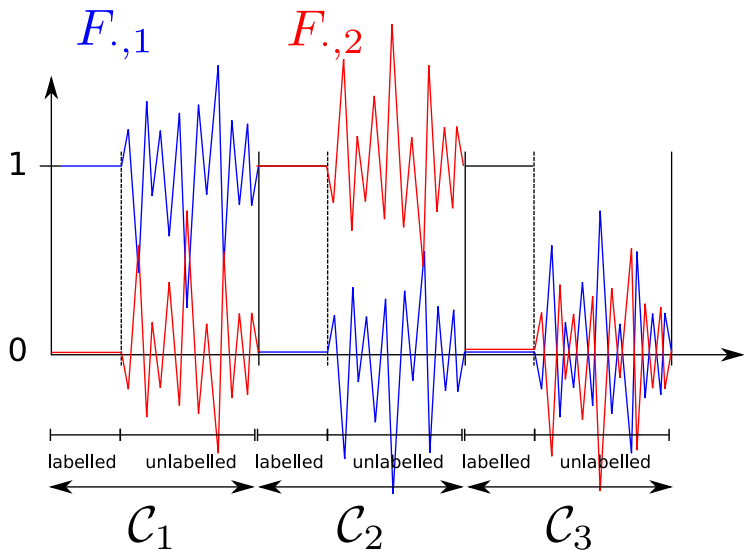


Figure: Typical expected performance output

## The finite-dimensional intuition: What we expect

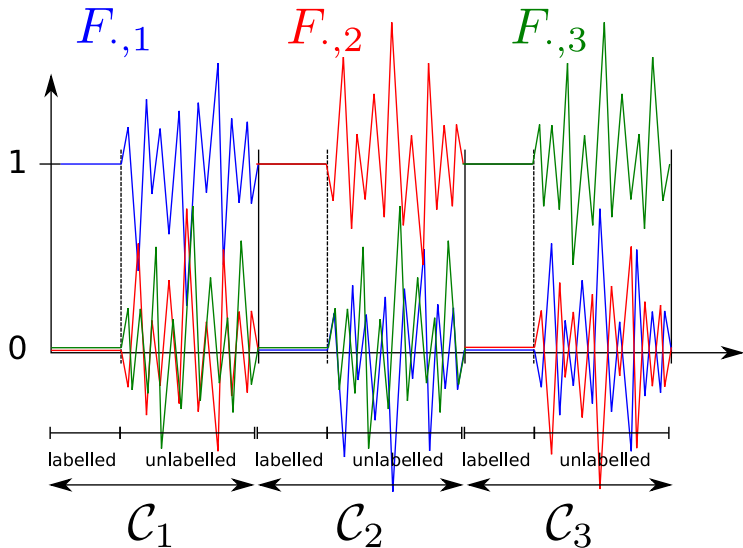


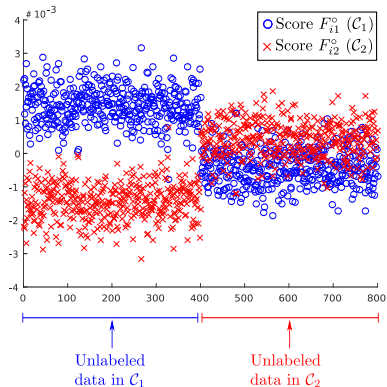
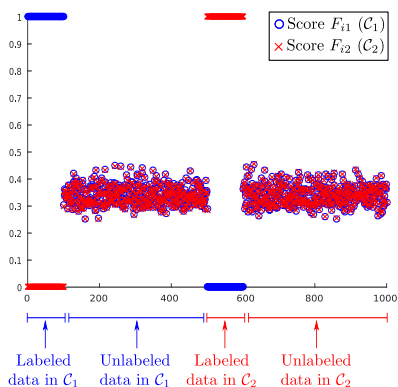
Figure: Typical expected performance output



## The reality: What we see!

**Setting.**  $p = 400$ ,  $n = 1000$ ,  $x_i \sim \mathcal{N}(\pm\mu, I_p)$ . Kernel  $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$ .

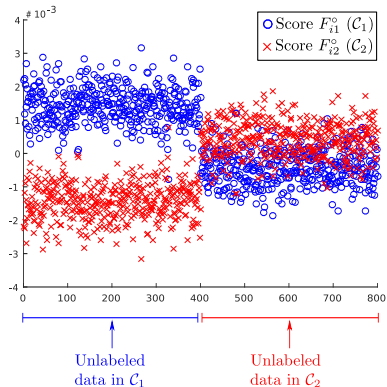
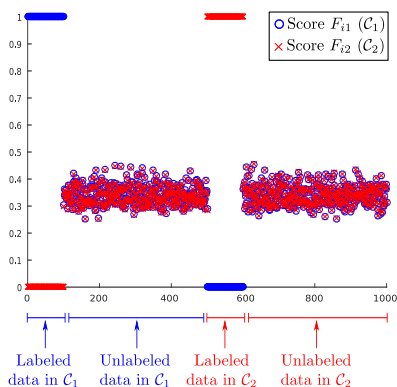
**Display.** Scores  $F_{ik}$  (left) and  $F_{ik} - \frac{1}{2}(F_{i1} + F_{i2})$  (right).



## The reality: What we see!

**Setting.**  $p = 400$ ,  $n = 1000$ ,  $x_i \sim \mathcal{N}(\pm\mu, I_p)$ . Kernel  $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$ .

**Display.** Scores  $F_{ik}$  (left) and  $F_{ik} - \frac{1}{2}(F_{i1} + F_{i2})$  (right).



⇒ Score are almost all identical... and do not follow the labelled data!

## MNIST Data Example

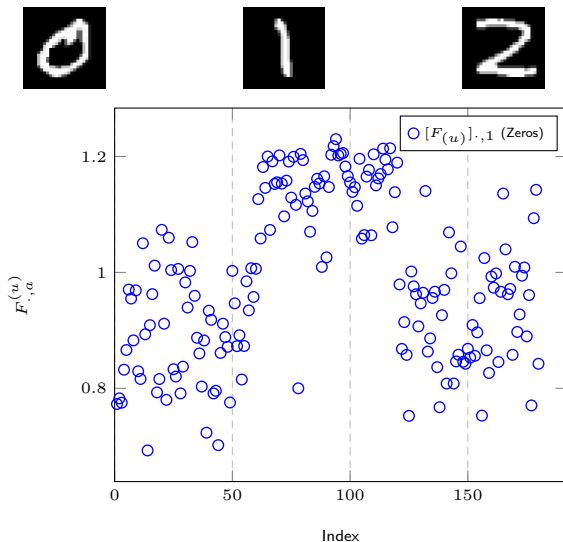


Figure: Vectors  $[F^{(u)}]_{:,a}$ ,  $a = 1, 2, 3$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example

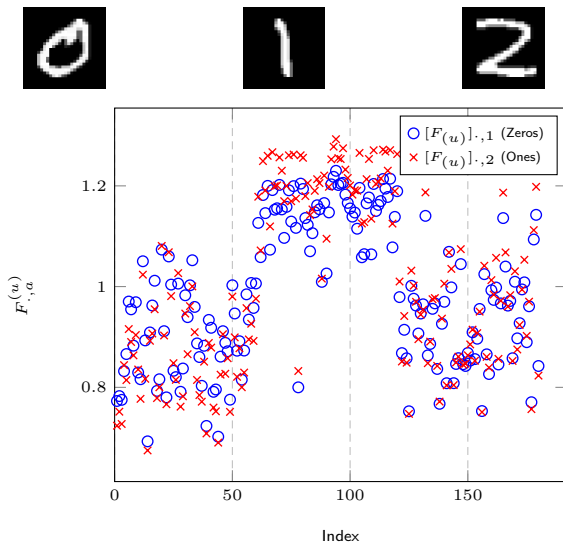


Figure: Vectors  $[F^{(u)}]_{\cdot, a}$ ,  $a = 1, 2, 3$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example

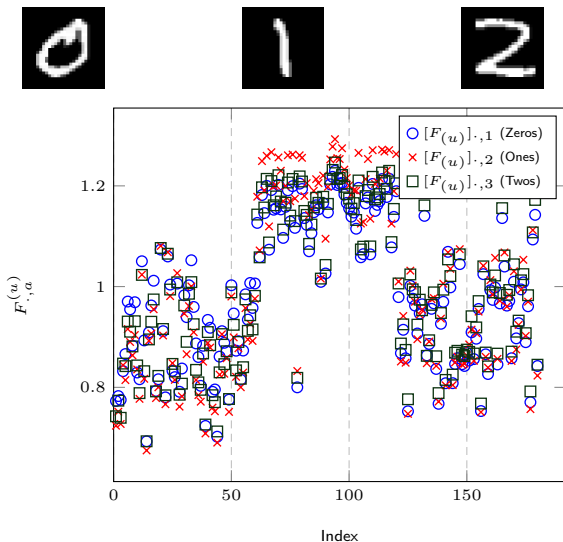


Figure: Vectors  $[F^{(u)}]_{\cdot, a}$ ,  $a = 1, 2, 3$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## Theoretical Findings

**Method:** Assume  $n_l/n \rightarrow c_l \in (0, 1)$

- ▶ We aim at characterizing

$$F^{(u)} = \left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

## Theoretical Findings

**Method:** Assume  $n_l/n \rightarrow c_l \in (0, 1)$

- ▶ We aim at characterizing

$$F^{(u)} = \left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

- ▶ Taylor expansion of  $K$  as  $n, p \rightarrow \infty$ ,

$$\begin{aligned} K_{(u,u)} &= f(\tau) 1_{n_u} 1_{n_u}^\top + O_{\|\cdot\|}(n^{-\frac{1}{2}}) \\ D_{(u)} &= n f(\tau) I_{n_u} + O(n^{\frac{1}{2}}) \end{aligned}$$

and similarly for  $K_{(u,l)}$ ,  $D_{(l)}$ .

## Theoretical Findings

**Method:** Assume  $n_l/n \rightarrow c_l \in (0, 1)$

- ▶ We aim at characterizing

$$F^{(u)} = \left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

- ▶ Taylor expansion of  $K$  as  $n, p \rightarrow \infty$ ,

$$\begin{aligned} K_{(u,u)} &= f(\tau) \mathbf{1}_{n_u} \mathbf{1}_{n_u}^\top + O_{\|\cdot\|}(n^{-\frac{1}{2}}) \\ D_{(u)} &= n f(\tau) I_{n_u} + O(n^{\frac{1}{2}}) \end{aligned}$$

and similarly for  $K_{(u,l)}$ ,  $D_{(l)}$ .

- ▶ So that

$$\left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} = \left( I_{n_u} - \frac{\mathbf{1}_{n_u} \mathbf{1}_{n_u}^\top}{n} + O_{\|\cdot\|}(n^{-\frac{1}{2}}) \right)^{-1}$$

easily Taylor expanded.



# Main Results

**Results:** Assuming  $n_l/n \rightarrow c_l \in (0, 1)$ , by previous Taylor expansion,

► In the first order,

$$F_{\cdot,a}^{(u)} = C \frac{n_{l,a}}{n} \left[ \underbrace{v}_{O(1)} + \underbrace{\alpha \frac{t_a \mathbf{1}_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})} \right] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

where  $v = O(1)$  random vector (entry-wise) and  $t_a = \frac{1}{\sqrt{p}} \text{tr } C_a^\circ$ .

# Main Results

**Results:** Assuming  $n_l/n \rightarrow c_l \in (0, 1)$ , by previous Taylor expansion,

► In the first order,

$$F_{\cdot,a}^{(u)} = C \frac{n_{l,a}}{n} \left[ \underbrace{v}_{O(1)} + \underbrace{\alpha \frac{t_a \mathbf{1}_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})} \right] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

where  $v = O(1)$  random vector (entry-wise) and  $t_a = \frac{1}{\sqrt{p}} \text{tr } C_a^\circ$ .

► Consequences:

# Main Results

**Results:** Assuming  $n_l/n \rightarrow c_l \in (0, 1)$ , by previous Taylor expansion,

- ▶ In the first order,

$$F_{\cdot,a}^{(u)} = C \frac{n_{l,a}}{n} \left[ \underbrace{v}_{O(1)} + \underbrace{\alpha \frac{t_a \mathbf{1}_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})} \right] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

where  $v = O(1)$  random vector (entry-wise) and  $t_a = \frac{1}{\sqrt{p}} \text{tr } C_a^o$ .

- ▶ Consequences:
  - ▶ Random non-informative bias  $v$

# Main Results

**Results:** Assuming  $n_l/n \rightarrow c_l \in (0, 1)$ , by previous Taylor expansion,

- ▶ In the first order,

$$F_{\cdot,a}^{(u)} = C \frac{n_{l,a}}{n} \left[ \underbrace{v}_{O(1)} + \underbrace{\alpha \frac{t_a \mathbf{1}_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})} \right] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

where  $v = O(1)$  random vector (entry-wise) and  $t_a = \frac{1}{\sqrt{p}} \text{tr } C_a^o$ .

- ▶ Consequences:
  - ▶ Random non-informative bias  $v$
  - ▶ Strong Impact of  $n_{l,a}$

$F_{\cdot,a}^{(u)}$  to be scaled by  $n_{l,a}$

# Main Results

**Results:** Assuming  $n_l/n \rightarrow c_l \in (0, 1)$ , by previous Taylor expansion,

- ▶ In the first order,

$$F_{\cdot,a}^{(u)} = C \frac{n_{l,a}}{n} \left[ \underbrace{v}_{O(1)} + \underbrace{\alpha \frac{t_a \mathbf{1}_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})} \right] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

where  $v = O(1)$  random vector (entry-wise) and  $t_a = \frac{1}{\sqrt{p}} \text{tr } C_a^O$ .

- ▶ Consequences:

- ▶ Random non-informative bias  $v$
- ▶ Strong Impact of  $n_{l,a}$

$$F_{\cdot,a}^{(u)} \text{ to be scaled by } n_{l,a}$$

- ▶ Additional per-class bias  $\alpha t_a \mathbf{1}_{n_u}$

$$\alpha = 0 + \frac{\beta}{\sqrt{p}}.$$

## Main Results

As a consequence of the remarks above, we take

$$\alpha = \frac{\beta}{\sqrt{p}}$$

and define

$$\hat{F}_{i,a}^{(u)} = \frac{np}{n_{l,a}} F_{ia}^{(u)}.$$

## Main Results

As a consequence of the remarks above, we take

$$\alpha = \frac{\beta}{\sqrt{p}}$$

and define

$$\hat{F}_{i,a}^{(u)} = \frac{np}{n_{l,a}} F_{ia}^{(u)}.$$

### Theorem

For  $x_i \in \mathcal{C}_b$  unlabelled,

$$\hat{F}_{i,\cdot} - G_b \rightarrow 0, \quad G_b \sim \mathcal{N}(m_b, \Sigma_b)$$

where  $m_b \in \mathbb{R}^k$ ,  $\Sigma_b \in \mathbb{R}^{k \times k}$  given by

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)} \tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)} \tilde{t}_a \tilde{t}_b + \frac{2f''(\tau)}{f(\tau)} \tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b + \beta \frac{n}{n_l} \frac{f'(\tau)}{f(\tau)} t_a + B_b$$
$$(\Sigma_b)_{a_1 a_2} = \frac{2 \operatorname{tr} C_b^2}{p} \left( \frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)} \right)^2 t_{a_1} t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2} \left( [M^T C_b M]_{a_1 a_2} + \frac{\delta_{a_1}^{a_2} p}{n_{l, a_1}} T_{b a_1} \right)$$

with  $t, T, M$  as before,  $\tilde{X}_a = X_a - \sum_{d=1}^k \frac{n_{l,d}}{n_l} X_d^\circ$  and  $B_b$  bias independent of  $a$ .

## Corollary (Asymptotic Classification Error)

For  $k = 2$  classes and  $a \neq b$ ,

$$P(\hat{F}_{i,a} > \hat{F}_{ib} \mid x_i \in \mathcal{C}_b) - Q\left(\frac{(m_b)_b - (m_b)_a}{\sqrt{[1, -1]\Sigma_b[1, -1]^T}}\right) \rightarrow 0.$$



## Corollary (Asymptotic Classification Error)

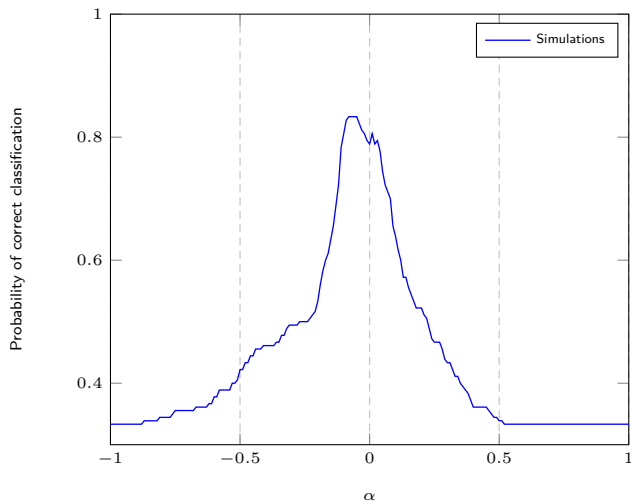
For  $k = 2$  classes and  $a \neq b$ ,

$$P(\hat{F}_{i,a} > \hat{F}_{ib} \mid x_i \in \mathcal{C}_b) - Q\left(\frac{(m_b)_b - (m_b)_a}{\sqrt{[1, -1]\Sigma_b[1, -1]^T}}\right) \rightarrow 0.$$

**Some consequences:**

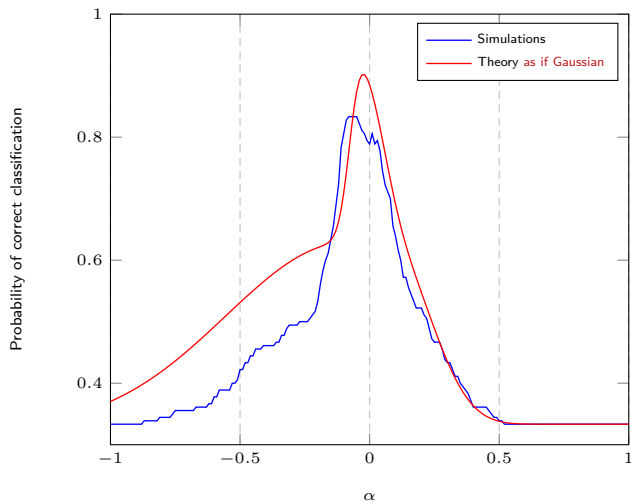
- ▶ non obvious choices of appropriate kernels
- ▶ non obvious choice of optimal  $\beta$  (induces a possibly beneficial bias)
- ▶ importance of  $n_l$  versus  $n_u$ .

## MNIST Data Example



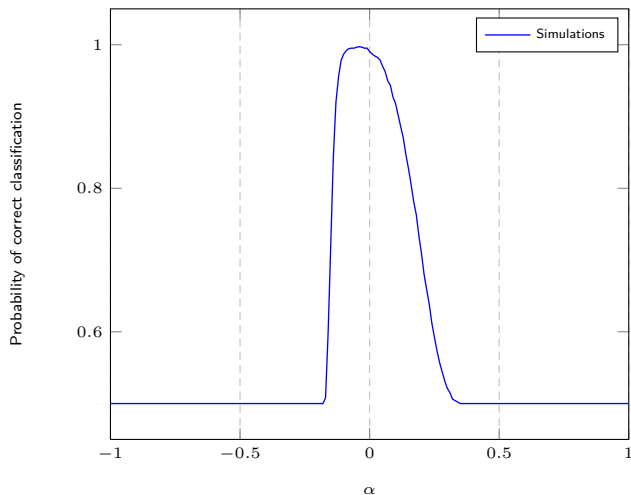
**Figure:** Performance as a function of  $\alpha$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example



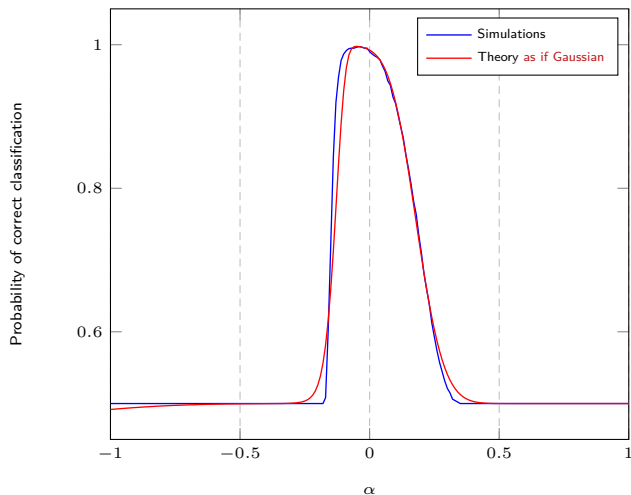
**Figure:** Performance as a function of  $\alpha$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example



**Figure:** Performance as a function of  $\alpha$ , for 2-class MNIST data (zeros, ones),  $n = 1568$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example



**Figure:** Performance as a function of  $\alpha$ , for 2-class MNIST data (zeros, ones),  $n = 1568$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## Is semi-supervised learning really semi-supervised?

### Reminder:

For  $x_i \in \mathcal{C}_b$  unlabelled,  $\hat{F}_{i,\cdot} - G_b \rightarrow 0$ ,  $G_b \sim \mathcal{N}(m_b, \Sigma_b)$  with

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)}\tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)}\tilde{t}_a\tilde{t}_b + \frac{2f''(\tau)}{f(\tau)}\tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2}t_a t_b + \beta \frac{n}{n_l} \frac{f'(\tau)}{f(\tau)} t_a + B_b$$

$$(\Sigma_b)_{a_1 a_2} = \frac{2\text{tr} C_b^2}{p} \left( \frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)} \right)^2 t_{a_1} t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2} \left( [M^\top C_b M]_{a_1 a_2} + \frac{\delta_{a_1}^{a_2} p}{n_{l, a_1}} T_{ba_1} \right)$$

with  $t, T, M$  as before,  $\tilde{X}_a = X_a - \sum_{d=1}^k \frac{n_{l,d}}{n_l} X_d^\circ$  and  $B_b$  bias independent of  $a$ .

## Is semi-supervised learning really semi-supervised?

### Reminder:

For  $x_i \in \mathcal{C}_b$  unlabelled,  $\hat{F}_{i,\cdot} - G_b \rightarrow 0$ ,  $G_b \sim \mathcal{N}(m_b, \Sigma_b)$  with

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)}\tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)}\tilde{t}_a\tilde{t}_b + \frac{2f''(\tau)}{f(\tau)}\tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2}t_a t_b + \beta \frac{n}{n_l} \frac{f'(\tau)}{f(\tau)} t_a + B_b$$

$$(\Sigma_b)_{a_1 a_2} = \frac{2\text{tr} C_b^2}{p} \left( \frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)} \right)^2 t_{a_1} t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2} \left( [M^\top C_b M]_{a_1 a_2} + \frac{\delta_{a_1}^{a_2} p}{n_{l,a_1}} T_{ba_1} \right)$$

with  $t, T, M$  as before,  $\tilde{X}_a = X_a - \sum_{d=1}^k \frac{n_{l,d}}{n_l} X_d^\circ$  and  $B_b$  bias independent of  $a$ .

### The problem with unlabelled data:

- ▶ Result **does not** depend on  $n_u$ !  
→ increasing  $n_u$  asymptotically non beneficial.

## Is semi-supervised learning really semi-supervised?

### Reminder:

For  $x_i \in \mathcal{C}_b$  unlabelled,  $\hat{F}_{i,\cdot} - G_b \rightarrow 0$ ,  $G_b \sim \mathcal{N}(m_b, \Sigma_b)$  with

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)}\tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)}\tilde{t}_a\tilde{t}_b + \frac{2f''(\tau)}{f(\tau)}\tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2}t_a t_b + \beta \frac{n}{n_l} \frac{f'(\tau)}{f(\tau)}t_a + B_b$$
$$(\Sigma_b)_{a_1 a_2} = \frac{2\text{tr} C_b^2}{p} \left( \frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)} \right)^2 t_{a_1} t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2} \left( [M^\top C_b M]_{a_1 a_2} + \frac{\delta_{a_1}^{a_2} p}{n_{l,a_1}} T_{ba_1} \right)$$

with  $t, T, M$  as before,  $\tilde{X}_a = X_a - \sum_{d=1}^k \frac{n_{l,d}}{n_l} X_d^\circ$  and  $B_b$  bias independent of  $a$ .

### The problem with unlabelled data:

- ▶ Result **does not** depend on  $n_u$ !  
→ increasing  $n_u$  asymptotically non beneficial.
- ▶ Even best Laplacian regularizer **brings SSL to be merely supervised learning**.



## Consequences of the finite-dimensional “mismatch”

## Consequences of the finite-dimensional “mismatch”

- ▶ A priori, **the algorithm should not work**

## Consequences of the finite-dimensional “mismatch”

- ▶ A priori, **the algorithm should not work**
- ▶ Indeed “in general” it does not!

## Consequences of the finite-dimensional “mismatch”

- ▶ A priori, **the algorithm should not work**
- ▶ Indeed “in general” it does not!
- ▶ But, luckily, after some (not clearly motivated) renormalization, it works again...

### Consequences of the finite-dimensional “mismatch”

- ▶ A priori, **the algorithm should not work**
- ▶ Indeed “in general” it does not!
- ▶ But, luckily, after some (not clearly motivated) renormalization, it works again...
  
- ▶ **BUT** it does not use efficiently unlabelled data!

## Consequences of the finite-dimensional “mismatch”

- ▶ A priori, **the algorithm should not work**
- ▶ Indeed “in general” it does not!
- ▶ But, luckily, after some (not clearly motivated) renormalization, it works again...
  
- ▶ **BUT** it does not use efficiently unlabelled data!

Chapelle, Schölkopf, Zien, “**Semi-Supervised Learning**”, Chapter 4, 2009.

*Our concern is this: it is frequently the case that we would be better off just discarding the unlabeled data and employing a supervised method, rather than taking a semi-supervised route. Thus we worry about the embarrassing situation where the addition of unlabeled data degrades the performance of a classifier.*

## Consequences of the finite-dimensional “mismatch”

- ▶ A priori, **the algorithm should not work**
- ▶ Indeed “in general” it does not!
- ▶ But, luckily, after some (not clearly motivated) renormalization, it works again...
  
- ▶ **BUT** it does not use efficiently unlabelled data!

Chapelle, Schölkopf, Zien, “**Semi-Supervised Learning**”, Chapter 4, 2009.

*Our concern is this: it is frequently the case that we would be better off just discarding the unlabeled data and employing a supervised method, rather than taking a semi-supervised route. Thus we worry about the embarrassing situation where the addition of unlabeled data degrades the performance of a classifier.*

What RMT can do about it

## Consequences of the finite-dimensional “mismatch”

- ▶ A priori, **the algorithm should not work**
- ▶ Indeed “in general” it does not!
- ▶ But, luckily, after some (not clearly motivated) renormalization, it works again...
  
- ▶ **BUT** it does not use efficiently unlabelled data!

Chapelle, Schölkopf, Zien, “**Semi-Supervised Learning**”, Chapter 4, 2009.

*Our concern is this: it is frequently the case that we would be better off just discarding the unlabeled data and employing a supervised method, rather than taking a semi-supervised route. Thus we worry about the embarrassing situation where the addition of unlabeled data degrades the performance of a classifier.*

## What RMT can do about it

- ▶ Asymptotic performance analysis: **clear understanding of what we see!**



# Exploiting RMT to resurrect SSL

## Consequences of the finite-dimensional “mismatch”

- ▶ A priori, **the algorithm should not work**
- ▶ Indeed “in general” it does not!
- ▶ But, luckily, after some (not clearly motivated) renormalization, it works again...
  
- ▶ **BUT** it does not use efficiently unlabelled data!

Chapelle, Schölkopf, Zien, “**Semi-Supervised Learning**”, Chapter 4, 2009.

*Our concern is this: it is frequently the case that we would be better off just discarding the unlabeled data and employing a supervised method, rather than taking a semi-supervised route. Thus we worry about the embarrassing situation where the addition of unlabeled data degrades the performance of a classifier.*

## What RMT can do about it

- ▶ Asymptotic performance analysis: **clear understanding of what we see!**
- ▶ Update the algorithm and **provably improve unlabelled data use.**

## Resurrecting SSL by centering (SSL Improved)

Reminder:

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^k \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2 \quad \text{with } F_{ia}^{(l)} = \delta_{\{x_i \in C_a\}}$$

$$\Leftrightarrow F^{(u)} = \left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}.$$

## Resurrecting SSL by centering (SSL Improved)

Reminder:

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^k \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2 \quad \text{with } F_{ia}^{(l)} = \delta_{\{x_i \in C_a\}}$$

$$\Leftrightarrow F^{(u)} = \left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}.$$

Domination of score flattening:

- ▶ **Consequence of**  $\frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_i\|^2 \rightarrow \tau$ :  $D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \simeq \frac{1}{n} \mathbf{1}_{n_u} \mathbf{1}_{n_u}^T$  and **clustering information vanishes** (not so obvious but can be shown).

# Resurrecting SSL by centering (SSL Improved)

Reminder:

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^k \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2 \quad \text{with } F_{ia}^{(l)} = \delta_{\{x_i \in C_a\}}$$

$$\Leftrightarrow F^{(u)} = \left( I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}.$$

Domination of score flattening:

- ▶ Consequence of  $\frac{1}{p} \|\mathbf{x}_i - \mathbf{x}_i\|^2 \rightarrow \tau$ :  $D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \simeq \frac{1}{n} \mathbf{1}_{n_u} \mathbf{1}_{n_u}^\top$  and **clustering information vanishes** (not so obvious but can be shown).

Solution:

- ▶ Forgetting finite-dimensional intuition: **“recenter”  $K$  to kill flattening**, i.e., use

$$\boxed{\tilde{K} = PKP}, \quad P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

## Asymptotic Performance Analysis

### Theorem ([Mai,C'19] Asymptotic Performance of Improved SSL)

For  $x_i \in \mathcal{C}_b$  unlabelled, score vector  $\hat{F}_{i,\cdot} \in \mathbb{R}^k$  with  $\tilde{K}$  satisfies:

$$\hat{F}_{i,\cdot} - \tilde{G}_b \rightarrow 0, \quad \tilde{G}_b \sim \mathcal{N}(\tilde{m}_b, \tilde{\Sigma}_b)$$

with  $\tilde{m}_b \in \mathbb{R}^k$ ,  $\tilde{\Sigma}_b \in \mathbb{R}^{k \times k}$  still function of  $f(\tau), f'(\tau), f''(\tau), \mu_1, \dots, \mu_k, C_1, \dots, C_k$ .

## Asymptotic Performance Analysis

### Theorem ([Mai,C'19] Asymptotic Performance of Improved SSL)

For  $x_i \in \mathcal{C}_b$  unlabelled, score vector  $\hat{F}_{i,\cdot} \in \mathbb{R}^k$  with  $\tilde{K}$  satisfies:

$$\hat{F}_{i,\cdot} - \tilde{G}_b \rightarrow 0, \tilde{G}_b \sim \mathcal{N}(\tilde{m}_b, \tilde{\Sigma}_b)$$

with  $\tilde{m}_b \in \mathbb{R}^k$ ,  $\tilde{\Sigma}_b \in \mathbb{R}^{k \times k}$  still function of  $f(\tau), f'(\tau), f''(\tau), \mu_1, \dots, \mu_k, C_1, \dots, C_k$ .

**Most importantly:**  $\tilde{m}_b, \tilde{\Sigma}_b$  now dependent of  $n_u$  (number of unlabelled data).

# Asymptotic Performance Analysis

## Theorem ([Mai,C'19] Asymptotic Performance of Improved SSL)

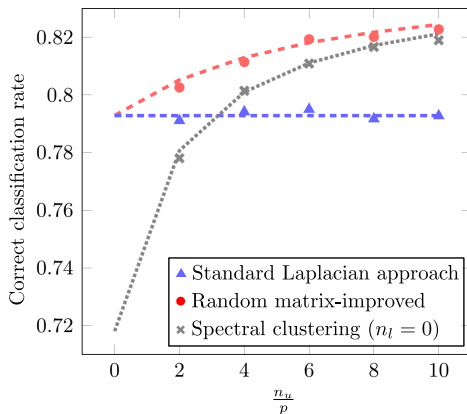
For  $x_i \in \mathcal{C}_b$  unlabelled, score vector  $\hat{F}_{i,\cdot} \in \mathbb{R}^k$  with  $\tilde{K}$  satisfies:

$$\hat{F}_{i,\cdot} - \tilde{G}_b \rightarrow 0, \quad \tilde{G}_b \sim \mathcal{N}(\tilde{m}_b, \tilde{\Sigma}_b)$$

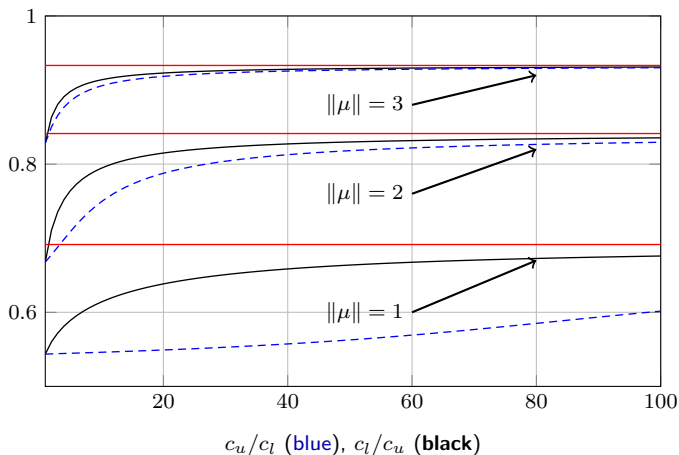
with  $\tilde{m}_b \in \mathbb{R}^k$ ,  $\tilde{\Sigma}_b \in \mathbb{R}^{k \times k}$  still function of  $f(\tau), f'(\tau), f''(\tau), \mu_1, \dots, \mu_k, C_1, \dots, C_k$ .

**Most importantly:**  $\tilde{m}_b, \tilde{\Sigma}_b$  now dependent of  $n_u$  (number of unlabelled data).

**Performances:**



## Performance as a function of $n_u$ , $n_l$ for $\mathcal{N}(\pm, I_p)$



**Figure:** Correct classification rate, at optimal  $\alpha$ , as a function of (i)  $n_u$  for fixed  $p/n_l = 5$  (blue) and (ii)  $n_l$  for fixed  $p/n_u = 5$  (black);  $c_1 = c_2 = \frac{1}{2}$ ; different values for  $\|\mu\|$ . Comparison to optimal Neyman–Pearson performance for known  $\mu$  (in red).



## Experimental evidence: MNIST



Digits	(0,8)	(2,7)	(6,9)
$n_u = 100$			
Centered kernel (RMT)	<b>89.5±3.6</b>	<b>89.5±3.4</b>	<b>85.3±5.9</b>
Iterated centered kernel (RMT)	<b>89.5±3.6</b>	<b>89.5±3.4</b>	<b>85.3±5.9</b>
Laplacian	75.5±5.6	74.2±5.8	70.0±5.5
Iterated Laplacian	87.2±4.7	86.0±5.2	81.4±6.8
Manifold	88.0±4.7	88.4±3.9	82.8±6.5
$n_u = 1000$			
Centered kernel (RMT)	92.2±0.9	92.5±0.8	92.6±1.6
Iterated centered kernel (RMT)	<b>92.3±0.9</b>	<b>92.5± 0.8</b>	<b>92.9±1.4</b>
Laplacian	65.6±4.1	74.4±4.0	69.5±3.7
Iterated Laplacian	<b>92.2±0.9</b>	92.4±0.9	92.0±1.6
Manifold	91.1±1.7	91.4±1.9	91.4±2.0

**Table:** Comparison of classification accuracy (%) on MNIST datasets with  $n_l = 10$ . Computed over 1000 random iterations for  $n_u = 100$  and 100 for  $n_u = 1000$ .

## Experimental evidence: Traffic signs (HOG features)



Class ID	(2,7)	(9,10)	(11,18)
$n_u = 100$			
Centered kernel (RMT)	79.0±10.4	77.5±9.2	78.5±7.1
Iterated centered kernel (RMT)	<b>85.3±5.9</b>	<b>89.2±5.6</b>	<b>90.1±6.7</b>
Laplacian	73.8±9.8	77.3±9.5	78.6±7.2
Iterated Laplacian	83.7±7.2	88.0±6.8	87.1±8.8
Manifold	77.6±8.9	81.4±10.4	82.3±10.8
$n_u = 1000$			
Centered kernel (RMT)	83.6±2.4	84.6±2.4	88.7±9.4
Iterated centered kernel (RMT)	<b>84.8±3.8</b>	<b>88.0±5.5</b>	<b>96.4±3.0</b>
Laplacian	72.7±4.2	88.9±5.7	95.8±3.2
Iterated Laplacian	83.0±5.5	88.2±6.0	92.7±6.1
Manifold	77.7±5.8	85.0±9.0	90.6±8.1

**Table:** Comparison of classification accuracy (%) on German Traffic Sign datasets with  $n_l = 10$ . Computed over 1000 random iterations for  $n_u = 100$  and 100 for  $n_u = 1000$ .

## Basics of Random Matrix Theory (**Romain COUILLET**)

- Motivation: Large Sample Covariance Matrices
- The Stieltjes Transform Method
- Spiked Models
- Other Common Random Matrix Models
- Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

- Motivation: EEG-based clustering*
- Covariance Distance Inference
- Revisiting Motivation*
- Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

- Support Vector Machines
- Semi-Supervised Learning
- From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)

## Notion of Concentrated Vectors

- ▶ **Observation:** RMT seems to predict ML performances for **real data** even with **Gaussian** assumptions!

---

<sup>2</sup>**Reminder:**  $\mathcal{F} : E \rightarrow F$  is  $\|\mathcal{F}\|_{lip}$ -Lipschitz if  $\forall (x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip} \|x - y\|_E$ .

## Notion of Concentrated Vectors

- ▶ **Observation:** RMT seems to predict ML performances for **real data** even with **Gaussian** assumptions!
- ▶ **But Real data** are **unlikely close** to **Gaussian**.

---

<sup>2</sup>**Reminder:**  $\mathcal{F} : E \rightarrow F$  is  $\|\mathcal{F}\|_{lip}$ -Lipschitz if  $\forall(x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip} \|x - y\|_E$ .

## Notion of Concentrated Vectors

- ▶ **Observation:** RMT seems to predict ML performances for **real data** even with **Gaussian** assumptions!
- ▶ **But Real data** are **unlikely close** to **Gaussian**.
- ▶ **Gaussian** vectors fall within a larger, more useful, class of random vectors!

---

<sup>2</sup>**Reminder:**  $\mathcal{F} : E \rightarrow F$  is  $\|\mathcal{F}\|_{lip}$ -Lipschitz if  $\forall(x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip} \|x - y\|_E$ .

# Notion of Concentrated Vectors

- ▶ **Observation:** RMT seems to predict ML performances for **real data** even with **Gaussian** assumptions!
- ▶ **But Real data** are **unlikely close** to **Gaussian**.
- ▶ **Gaussian** vectors fall within a larger, more useful, class of random vectors!

## Definition

Given a normed space  $(E, \|\cdot\|_E)$  and  $q \in \mathbb{R}$ , a random vector  $\mathbf{z} \in E$  is  $q$ -exponentially **concentrated** if for any 1-Lipschitz function<sup>2</sup>  $\mathcal{F} : \mathbb{R}^p \rightarrow \mathbb{R}$ , there exists  $C, c > 0$  s.t.

$$\mathbb{P}\{|\mathcal{F}(\mathbf{z}) - \mathbb{E}\mathcal{F}(\mathbf{z})| > t\} \leq Ce^{-ct^q}$$

---

<sup>2</sup>**Reminder:**  $\mathcal{F} : E \rightarrow F$  is  $\|\mathcal{F}\|_{lip}$ -Lipschitz if  $\forall(x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip} \|x - y\|_E$ .

# Notion of Concentrated Vectors

- ▶ **Observation:** RMT seems to predict ML performances for **real data** even with **Gaussian** assumptions!
- ▶ **But Real data** are **unlikely close** to **Gaussian**.
- ▶ **Gaussian** vectors fall within a larger, more useful, class of random vectors!

## Definition

Given a normed space  $(E, \|\cdot\|_E)$  and  $q \in \mathbb{R}$ , a random vector  $\mathbf{z} \in E$  is  $q$ -exponentially **concentrated** if for any 1-Lipschitz function<sup>2</sup>  $\mathcal{F} : \mathbb{R}^p \rightarrow \mathbb{R}$ , there exists  $C, c > 0$  s.t.

$$\mathbb{P}\{|\mathcal{F}(\mathbf{z}) - \mathbb{E}\mathcal{F}(\mathbf{z})| > t\} \leq C e^{-c t^q} \xrightarrow{\text{denoted}} \boxed{\mathbf{z} \in \mathcal{O}(e^{-\cdot^q})}$$

---

<sup>2</sup>Reminder:  $\mathcal{F} : E \rightarrow F$  is  $\|\mathcal{F}\|_{lip}$ -Lipschitz if  $\forall(x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip} \|x - y\|_E$ .



## Notion of Concentrated Vectors

- ▶ **Observation:** RMT seems to predict ML performances for **real data** even with **Gaussian** assumptions!
- ▶ **But Real data** are **unlikely close** to **Gaussian**.
- ▶ **Gaussian** vectors fall within a larger, more useful, class of random vectors!

### Definition

Given a normed space  $(E, \|\cdot\|_E)$  and  $q \in \mathbb{R}$ , a random vector  $\mathbf{z} \in E$  is  $q$ -exponentially **concentrated** if for any 1-Lipschitz function<sup>2</sup>  $\mathcal{F} : \mathbb{R}^p \rightarrow \mathbb{R}$ , there exists  $C, c > 0$  s.t.

$$\mathbb{P}\{|\mathcal{F}(\mathbf{z}) - \mathbb{E}\mathcal{F}(\mathbf{z})| > t\} \leq C e^{-c t^q} \xrightarrow{\text{denoted}} \boxed{\mathbf{z} \in \mathcal{O}(e^{-\cdot^q})}$$

(P1)  $X \sim \mathcal{N}(0, I_p)$  is 2-exponentially **concentrated**.

---

<sup>2</sup>Reminder:  $\mathcal{F} : E \rightarrow F$  is  $\|\mathcal{F}\|_{\text{Lip}}$ -Lipschitz if  $\forall(x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{\text{Lip}} \|x - y\|_E$ .

# Notion of Concentrated Vectors

- ▶ **Observation:** RMT seems to predict ML performances for **real data** even with **Gaussian** assumptions!
- ▶ **But Real data** are **unlikely close** to **Gaussian**.
- ▶ **Gaussian** vectors fall within a larger, more useful, class of random vectors!

## Definition

Given a normed space  $(E, \|\cdot\|_E)$  and  $q \in \mathbb{R}$ , a random vector  $\mathbf{z} \in E$  is  $q$ -exponentially **concentrated** if for any 1-Lipschitz function<sup>2</sup>  $\mathcal{F} : \mathbb{R}^p \rightarrow \mathbb{R}$ , there exists  $C, c > 0$  s.t.

$$\mathbb{P}\{|\mathcal{F}(\mathbf{z}) - \mathbb{E}\mathcal{F}(\mathbf{z})| > t\} \leq C e^{-c t^q} \xrightarrow{\text{denoted}} \boxed{\mathbf{z} \in \mathcal{O}(e^{-\cdot^q})}$$

**(P1)**  $X \sim \mathcal{N}(0, I_p)$  is 2-exponentially **concentrated**.

**(P2)** If  $X \in \mathcal{O}(e^{-\cdot^q})$  and  $\mathcal{G}$  is  $\|\mathcal{G}\|_{lip}$ -Lipschitz, then

$$\mathcal{G}(X) \in \mathcal{O}\left(e^{-(\cdot/\|\mathcal{G}\|_{lip})^q}\right).$$

---

<sup>2</sup>Reminder:  $\mathcal{F} : E \rightarrow F$  is  $\|\mathcal{F}\|_{lip}$ -Lipschitz if  $\forall (x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip} \|x - y\|_E$ .

# Notion of Concentrated Vectors

- ▶ **Observation:** RMT seems to predict ML performances for **real data** even with **Gaussian** assumptions!
- ▶ **But Real data** are **unlikely close** to **Gaussian**.
- ▶ **Gaussian** vectors fall within a larger, more useful, class of random vectors!

## Definition

Given a normed space  $(E, \|\cdot\|_E)$  and  $q \in \mathbb{R}$ , a random vector  $\mathbf{z} \in E$  is  $q$ -exponentially **concentrated** if for any 1-Lipschitz function<sup>2</sup>  $\mathcal{F} : \mathbb{R}^p \rightarrow \mathbb{R}$ , there exists  $C, c > 0$  s.t.

$$\mathbb{P}\{|\mathcal{F}(\mathbf{z}) - \mathbb{E}\mathcal{F}(\mathbf{z})| > t\} \leq Ce^{-ct^q} \xrightarrow{\text{denoted}} \boxed{\mathbf{z} \in \mathcal{O}(e^{-\cdot^q})}$$

(P1)  $X \sim \mathcal{N}(0, I_p)$  is 2-exponentially **concentrated**.

(P2) If  $X \in \mathcal{O}(e^{-\cdot^q})$  and  $\mathcal{G}$  is  $\|\mathcal{G}\|_{lip}$ -Lipschitz, then

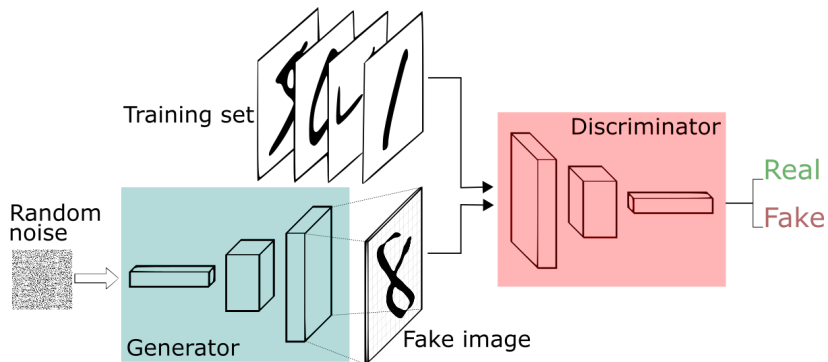
$$\mathcal{G}(X) \in \mathcal{O}\left(e^{-(\cdot/\|\mathcal{G}\|_{lip})^q}\right).$$

“Concentrated vectors are stable through Lipschitz maps.”

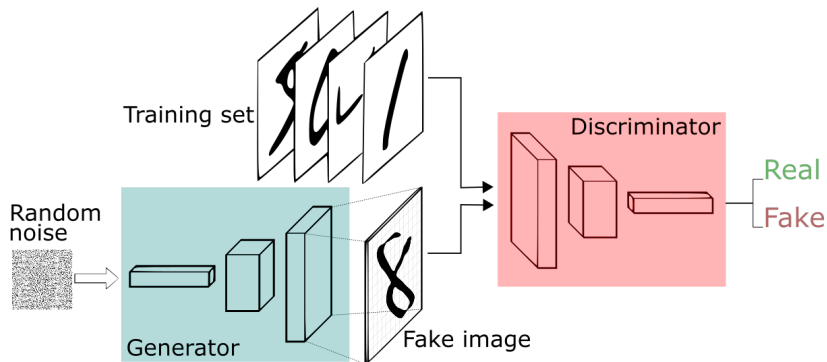
---

<sup>2</sup>Reminder:  $\mathcal{F} : E \rightarrow F$  is  $\|\mathcal{F}\|_{lip}$ -Lipschitz if  $\forall(x, y) \in E^2 : \|\mathcal{F}(x) - \mathcal{F}(y)\|_F \leq \|\mathcal{F}\|_{lip} \|x - y\|_E$ .

## GAN data: An Example of Concentrated Vectors



## GAN data: An Example of Concentrated Vectors



$$\min_G \max_D \mathbb{E}_{x \sim p(x)} [\log D(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D(G(z)))]$$

We generate data as

$$\text{Generated image} = G(\text{Gaussian})$$

## GAN data: An Example of Concentrated Vectors



Figure: Images generated by the BigGAN model [Brock *et al*, ICLR'19].

## GAN data: An Example of Concentrated Vectors



Figure: Images generated by the BigGAN model [Brock *et al*, ICLR'19].

$$\text{GAN Data} = \mathcal{F}_1 \circ \mathcal{F}_2 \circ \dots \circ \mathcal{F}_N(\text{Gaussian})$$

## GAN data: An Example of Concentrated Vectors



Figure: Images generated by the BigGAN model [Brock *et al*, ICLR'19].

$$\text{GAN Data} = \mathcal{F}_1 \circ \mathcal{F}_2 \circ \dots \circ \mathcal{F}_N(\text{Gaussian})$$

where the  $\mathcal{F}_i$ 's are either Fully Connected Layers, Convolutional Layers, Pooling Layers and Activation Functions, Residual Connections or Batch Normalizations.



## GAN data: An Example of Concentrated Vectors



Figure: Images generated by the BigGAN model [Brock *et al*, ICLR'19].

$$\text{GAN Data} = \mathcal{F}_1 \circ \mathcal{F}_2 \circ \dots \circ \mathcal{F}_N(\text{Gaussian})$$

where the  $\mathcal{F}_i$ 's are either Fully Connected Layers, Convolutional Layers, Pooling Layers and Activation Functions, Residual Connections or Batch Normalizations.

⇒ The  $\mathcal{F}_i$ 's are *Lipschitz* operations.

- ▶ **Fully Connected Layers and Convolutional Layers** are affine operations:

$$\mathcal{F}_i(x) = W_i x + b_i,$$

and  $\|\mathcal{F}_i\|_{lip} = \sup_{u \neq 0} \frac{\|W_i u\|_p}{\|u\|_p}$ , for any  $p$ -norm.

- ▶ **Fully Connected Layers and Convolutional Layers** are affine operations:

$$\mathcal{F}_i(x) = W_i x + b_i,$$

and  $\|\mathcal{F}_i\|_{lip} = \sup_{u \neq 0} \frac{\|W_i u\|_p}{\|u\|_p}$ , for any  $p$ -norm.

- ▶ **Pooling Layers and Activation Functions:** Are 1-Lipschitz operations with respect to any  $p$ -norm (e.g., ReLU and Max-pooling).

- ▶ **Fully Connected Layers and Convolutional Layers** are affine operations:

$$\mathcal{F}_i(x) = W_i x + b_i,$$

and  $\|\mathcal{F}_i\|_{lip} = \sup_{u \neq 0} \frac{\|W_i u\|_p}{\|u\|_p}$ , for any  $p$ -norm.

- ▶ **Pooling Layers and Activation Functions:** Are 1-Lipschitz operations with respect to any  $p$ -norm (e.g., ReLU and Max-pooling).

- ▶ **Residual Connections:**  $\mathcal{F}_i(x) = x + \mathcal{F}_i^{(1)} \circ \dots \circ \mathcal{F}_i^{(\ell)}(x)$

where the  $\mathcal{F}_i^{(j)}$ 's are Lipschitz operations, thus  $\mathcal{F}_i$  is a Lipschitz operation with Lipschitz constant bounded by  $1 + \prod_{j=1}^{\ell} \|\mathcal{F}_i^{(j)}\|_{lip}$ .

- ▶ ...

## Mixture of Concentrated Vectors

Consider data distributed in  $k$  classes  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$  as

$$X = \underbrace{[x_1, \dots, x_{n_1}]}_{\in \mathcal{O}(e^{-\cdot q_1})}, \underbrace{[x_{n_1+1}, \dots, x_{n_2}]}_{\in \mathcal{O}(e^{-\cdot q_2})}, \dots, \underbrace{[x_{n-n_k+1}, \dots, x_n]}_{\in \mathcal{O}(e^{-\cdot q_k})} \in \mathbb{R}^{p \times n}$$

## Mixture of Concentrated Vectors

Consider data distributed in  $k$  classes  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$  as

$$X = \underbrace{[x_1, \dots, x_{n_1}]}_{\in \mathcal{O}(e^{-\cdot q_1})}, \underbrace{[x_{n_1+1}, \dots, x_{n_2}]}_{\in \mathcal{O}(e^{-\cdot q_2})}, \dots, \underbrace{[x_{n-n_k+1}, \dots, x_n]}_{\in \mathcal{O}(e^{-\cdot q_k})} \in \mathbb{R}^{p \times n}$$

Denote

$$\mu_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i], \quad C_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i x_i^\top]$$

# Mixture of Concentrated Vectors

Consider data distributed in  $k$  classes  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$  as

$$X = \underbrace{[x_1, \dots, x_{n_1}]}_{\in \mathcal{O}(e^{-\cdot q_1})}, \underbrace{[x_{n_1+1}, \dots, x_{n_2}]}_{\in \mathcal{O}(e^{-\cdot q_2})}, \dots, \underbrace{[x_{n-n_k+1}, \dots, x_n]}_{\in \mathcal{O}(e^{-\cdot q_k})} \in \mathbb{R}^{p \times n}$$

Denote

$$\mu_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i], \quad C_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i x_i^\top]$$

Assumption (Growth rate)

As  $p \rightarrow \infty$ ,

# Mixture of Concentrated Vectors

Consider data distributed in  $k$  classes  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$  as

$$X = \underbrace{[x_1, \dots, x_{n_1}]}_{\in \mathcal{O}(e^{-\cdot q_1})}, \underbrace{[x_{n_1+1}, \dots, x_{n_2}]}_{\in \mathcal{O}(e^{-\cdot q_2})}, \dots, \underbrace{[x_{n-n_k+1}, \dots, x_n]}_{\in \mathcal{O}(e^{-\cdot q_k})} \in \mathbb{R}^{p \times n}$$

Denote

$$\mu_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i], \quad C_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i x_i^\top]$$

Assumption (Growth rate)

As  $p \rightarrow \infty$ ,

1.  $p/n \rightarrow c \in (0, \infty)$ .



# Mixture of Concentrated Vectors

Consider data distributed in  $k$  classes  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$  as

$$X = \underbrace{[x_1, \dots, x_{n_1}]}_{\in \mathcal{O}(e^{-\cdot q_1})}, \underbrace{[x_{n_1+1}, \dots, x_{n_2}]}_{\in \mathcal{O}(e^{-\cdot q_2})}, \dots, \underbrace{[x_{n-n_k+1}, \dots, x_n]}_{\in \mathcal{O}(e^{-\cdot q_k})} \in \mathbb{R}^{p \times n}$$

Denote

$$\mu_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i], \quad C_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i x_i^\top]$$

## Assumption (Growth rate)

As  $p \rightarrow \infty$ ,

1.  $p/n \rightarrow c \in (0, \infty)$ .
2. *The number of classes  $k$  is bounded.*

# Mixture of Concentrated Vectors

Consider data distributed in  $k$  classes  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$  as

$$X = \underbrace{[x_1, \dots, x_{n_1}]}_{\in \mathcal{O}(e^{-q_1})}, \underbrace{[x_{n_1+1}, \dots, x_{n_2}]}_{\in \mathcal{O}(e^{-q_2})}, \dots, \underbrace{[x_{n-n_k+1}, \dots, x_n]}_{\in \mathcal{O}(e^{-q_k})} \in \mathbb{R}^{p \times n}$$

Denote

$$\mu_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i], \quad C_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i x_i^\top]$$

## Assumption (Growth rate)

As  $p \rightarrow \infty$ ,

1.  $p/n \rightarrow c \in (0, \infty)$ .
2. The number of classes  $k$  is bounded.
3. For any  $\ell \in [k]$ ,  $\|\mu_\ell\| = \mathcal{O}(\sqrt{p})$ .

# Mixture of Concentrated Vectors

Consider data distributed in  $k$  classes  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k$  as

$$X = \underbrace{[x_1, \dots, x_{n_1}]}_{\in \mathcal{O}(e^{-\cdot q_1})}, \underbrace{[x_{n_1+1}, \dots, x_{n_2}]}_{\in \mathcal{O}(e^{-\cdot q_2})}, \dots, \underbrace{[x_{n-n_k+1}, \dots, x_n]}_{\in \mathcal{O}(e^{-\cdot q_k})} \in \mathbb{R}^{p \times n}$$

Denote

$$\mu_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i], \quad C_\ell = \mathbb{E}_{x_i \in \mathcal{C}_\ell} [x_i x_i^\top]$$

## Assumption (Growth rate)

As  $p \rightarrow \infty$ ,

1.  $p/n \rightarrow c \in (0, \infty)$ .
2. The number of classes  $k$  is bounded.
3. For any  $\ell \in [k]$ ,  $\|\mu_\ell\| = \mathcal{O}(\sqrt{p})$ .

## Notation

$$Q(z) = (X^\top X/p + zI_n)^{-1}.$$

## Behavior of Gram Matrices for Concentrated Vectors

### Theorem

Under the assumptions above, we have  $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\cdot)^q})$  in  $(\mathbb{R}^{n \times n}, \|\cdot\|)$ .  
Furthermore,

$$\|\mathbb{E}[Q(z)] - \tilde{Q}(z)\| = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \text{ where } \tilde{Q}(z) = \frac{1}{z}\Lambda(z) + \frac{1}{pz}J\Omega(z)J^\top$$

## Behavior of Gram Matrices for Concentrated Vectors

### Theorem

Under the assumptions above, we have  $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\cdot)^q})$  in  $(\mathbb{R}^{n \times n}, \|\cdot\|)$ .  
Furthermore,

$$\|\mathbb{E}[Q(z)] - \tilde{Q}(z)\| = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \text{ where } \tilde{Q}(z) = \frac{1}{z}\Lambda(z) + \frac{1}{pz}J\Omega(z)J^\top$$

with  $\Lambda(z) = \text{diag}\left\{\frac{1_{n_\ell}}{1+\delta_\ell(z)}\right\}_{\ell=1}^k$  and  $\Omega(z) = \text{diag}\{\mu_\ell^\top \tilde{R}(z)\mu_\ell\}_{\ell=1}^k$

## Behavior of Gram Matrices for Concentrated Vectors

### Theorem

Under the assumptions above, we have  $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\cdot)^q})$  in  $(\mathbb{R}^{n \times n}, \|\cdot\|)$ .  
Furthermore,

$$\|\mathbb{E}[Q(z)] - \tilde{Q}(z)\| = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \text{ where } \tilde{Q}(z) = \frac{1}{z}\Lambda(z) + \frac{1}{pz}J\Omega(z)J^\top$$

with  $\Lambda(z) = \text{diag}\left\{\frac{1_{n_\ell}}{1+\delta_\ell(z)}\right\}_{\ell=1}^k$  and  $\Omega(z) = \text{diag}\{\mu_\ell^\top \tilde{R}(z) \mu_\ell\}_{\ell=1}^k$

$$\tilde{R}(z) = \left(\frac{1}{k} \sum_{\ell=1}^k \frac{C_\ell}{1+\delta_\ell(z)} + zI_p\right)^{-1}$$

## Behavior of Gram Matrices for Concentrated Vectors

### Theorem

Under the assumptions above, we have  $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\cdot)^q})$  in  $(\mathbb{R}^{n \times n}, \|\cdot\|)$ .  
Furthermore,

$$\|\mathbb{E}[Q(z)] - \tilde{Q}(z)\| = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \text{ where } \tilde{Q}(z) = \frac{1}{z}\Lambda(z) + \frac{1}{pz}J\Omega(z)J^\top$$

with  $\Lambda(z) = \text{diag}\left\{\frac{1_{n_\ell}}{1+\delta_\ell(z)}\right\}_{\ell=1}^k$  and  $\Omega(z) = \text{diag}\{\mu_\ell^\top \tilde{R}(z) \mu_\ell\}_{\ell=1}^k$

$$\tilde{R}(z) = \left(\frac{1}{k} \sum_{\ell=1}^k \frac{C_\ell}{1+\delta_\ell(z)} + zI_p\right)^{-1}$$

with  $\delta(z) = [\delta_1(z), \dots, \delta_k(z)]$  is the unique fixed point of the system of equations

$$\delta_\ell(z) = \frac{1}{p} \text{tr} \left( C_\ell \left( \frac{1}{k} \sum_{j=1}^k \frac{C_j}{1+\delta_j(z)} + zI_p \right)^{-1} \right) \text{ for each } \ell \in [k].$$

# Behavior of Gram Matrices for Concentrated Vectors

## Theorem

Under the assumptions above, we have  $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\cdot)^q})$  in  $(\mathbb{R}^{n \times n}, \|\cdot\|)$ .  
Furthermore,

$$\|\mathbb{E}[Q(z)] - \tilde{Q}(z)\| = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \text{ where } \tilde{R}(z) = \frac{1}{z}\Lambda(z) + \frac{1}{pz}J\Omega(z)J^\top$$

with  $\Lambda(z) = \text{diag}\left\{\frac{1_{n_\ell}}{1+\delta_\ell(z)}\right\}_{\ell=1}^k$  and  $\Omega(z) = \text{diag}\{\mu_\ell^\top \tilde{R}(z) \mu_\ell\}_{\ell=1}^k$

$$\tilde{R}(z) = \left(\frac{1}{k} \sum_{\ell=1}^k \frac{C_\ell}{1 + \delta_\ell(z)} + zI_p\right)^{-1}$$

with  $\delta(z) = [\delta_1(z), \dots, \delta_k(z)]$  is the unique fixed point of the system of equations

$$\delta_\ell(z) = \text{tr}\left(C_\ell \left(\frac{1}{k} \sum_{j=1}^k \frac{C_j}{1 + \delta_j(z)} + zI_p\right)^{-1}\right) \text{ for each } \ell \in [k].$$



# Behavior of Gram Matrices for Concentrated Vectors

## Theorem

Under the assumptions above, we have  $Q(z) \in \mathcal{O}(e^{-(\sqrt{p}\cdot)^q})$  in  $(\mathbb{R}^{n \times n}, \|\cdot\|)$ .  
Furthermore,

$$\|\mathbb{E}[Q(z)] - \tilde{Q}(z)\| = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right) \text{ where } \tilde{R}(z) = \frac{1}{z}\Lambda(z) + \frac{1}{pz}J\Omega(z)J^\top$$

with  $\Lambda(z) = \text{diag}\left\{\frac{1_{n_\ell}}{1+\delta_\ell(z)}\right\}_{\ell=1}^k$  and  $\Omega(z) = \text{diag}\{\mu_\ell^\top \tilde{R}(z) \mu_\ell\}_{\ell=1}^k$

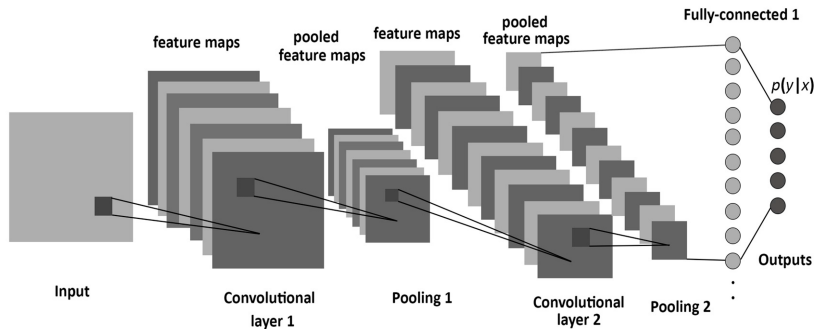
$$\tilde{R}(z) = \left(\frac{1}{k} \sum_{\ell=1}^k \frac{C_\ell}{1 + \delta_\ell(z)} + zI_p\right)^{-1}$$

with  $\delta(z) = [\delta_1(z), \dots, \delta_k(z)]$  is the unique fixed point of the system of equations

$$\delta_\ell(z) = \text{tr}\left(C_\ell \left(\frac{1}{k} \sum_{j=1}^k \frac{C_j}{1 + \delta_j(z)} + zI_p\right)^{-1}\right) \text{ for each } \ell \in [k].$$

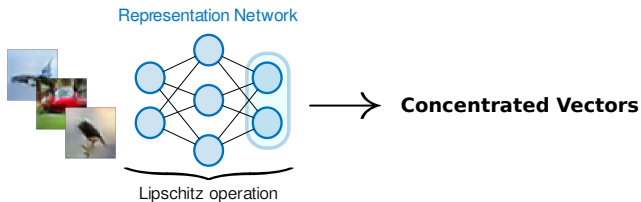
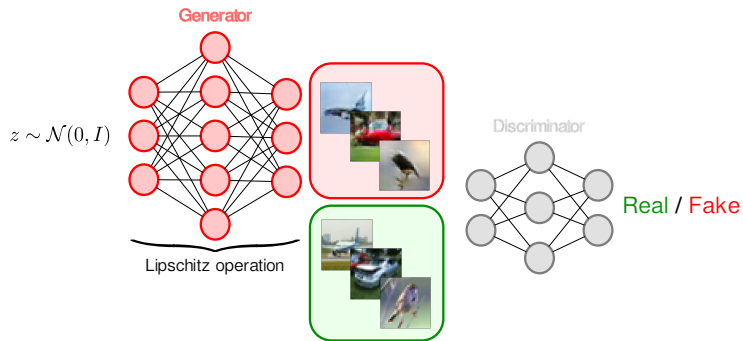
**Key Observation:** Only first and second order statistics matter!

# Application to CNN Representations of GAN Images



- ▶ CNN representations correspond to the **one before last** layer.

# Application to CNN Representations of GAN Images

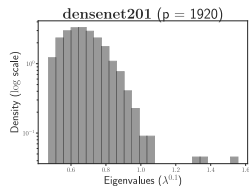
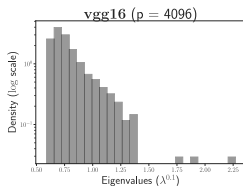
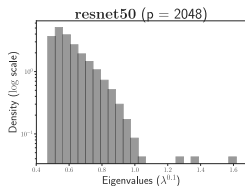


# Application to CNN representations of GAN Images

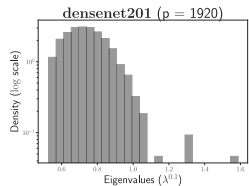
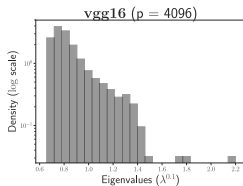
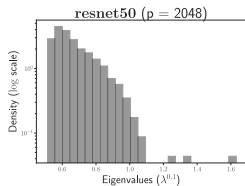


# Application to CNN representations of GAN Images

GAN Images

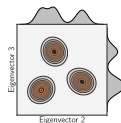
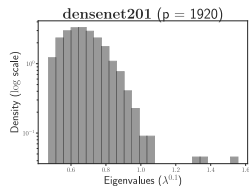
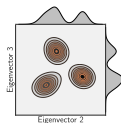
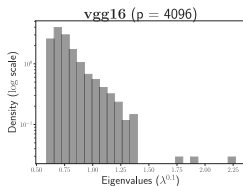
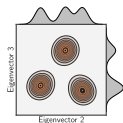
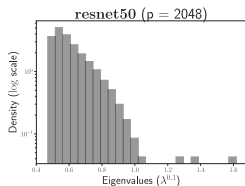


Real Images

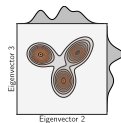
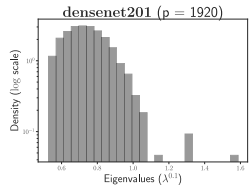
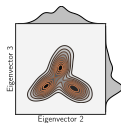
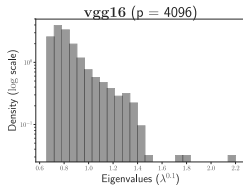
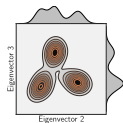
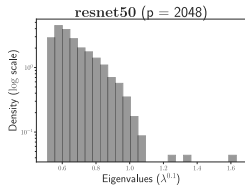


# Application to CNN representations of GAN Images

## GAN Images

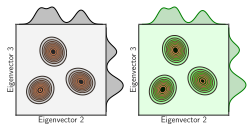
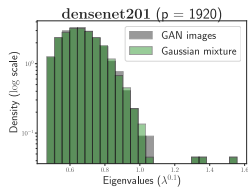
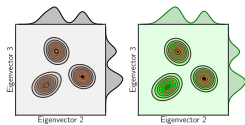
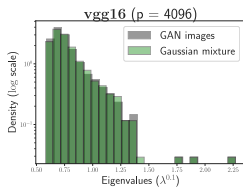
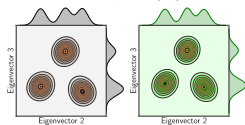
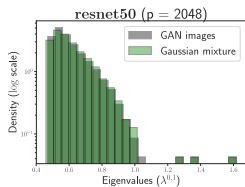


## Real Images

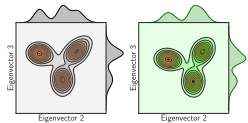
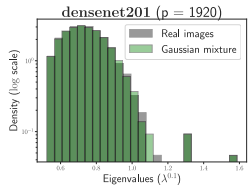
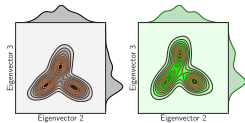
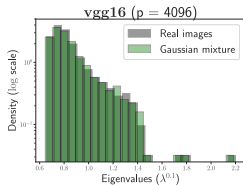
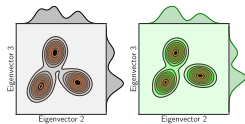
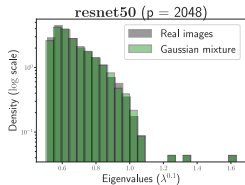


# Application to CNN representations of GAN Images

## GAN Images



## Real Images



## Basics of Random Matrix Theory (**Romain COUILLET**)

- Motivation: Large Sample Covariance Matrices

- The Stieltjes Transform Method

- Spiked Models

- Other Common Random Matrix Models

- Applications

## Large dimensional inference and kernels (**Malik TIOMOKO**)

- Motivation: EEG-based clustering*

- Covariance Distance Inference

- Revisiting Motivation*

- Kernel Asymptotics

## Application to machine learning (**Mohamed SEDDIK**)

- Support Vector Machines

- Semi-Supervised Learning

- From Gaussian Mixtures to Real Data

## Take-away Messages, Summary of Results and Perspectives (**Romain COUILLET**)



## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p$

## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p \Rightarrow$  **simplification in analyses and models.**

## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p \Rightarrow$  **simplification in analyses and models.**
- ▶ Non-trivial **phase transition** phenomena (ability to detect, estimate) when  $p, n \rightarrow \infty$ .

## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p \Rightarrow$  **simplification in analyses and models.**
- ▶ Non-trivial **phase transition** phenomena (ability to detect, estimate) when  $p, n \rightarrow \infty$ .
- ▶ Access to **limiting performances** and not only bounds!

## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p \Rightarrow$  **simplification in analyses and models.**
- ▶ Non-trivial **phase transition** phenomena (ability to detect, estimate) when  $p, n \rightarrow \infty$ .
- ▶ Access to **limiting performances** and not only bounds!  $\Rightarrow$  **hyperparameter optimization, algorithm improvement.**

## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p \Rightarrow$  **simplification in analyses and models.**
- ▶ Non-trivial **phase transition** phenomena (ability to detect, estimate) when  $p, n \rightarrow \infty$ .
- ▶ Access to **limiting performances** and not only bounds!  $\Rightarrow$  **hyperparameter optimization, algorithm improvement.**
- ▶ **Complete intuitive change**

## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p \Rightarrow$  **simplification in analyses and models.**
- ▶ Non-trivial **phase transition** phenomena (ability to detect, estimate) when  $p, n \rightarrow \infty$ .
- ▶ Access to **limiting performances** and not only bounds!  $\Rightarrow$  **hyperparameter optimization, algorithm improvement.**
- ▶ **Complete intuitive change**  $\Rightarrow$  **opens way to renewed methods.**

## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p \Rightarrow$  **simplification in analyses and models.**
- ▶ Non-trivial **phase transition** phenomena (ability to detect, estimate) when  $p, n \rightarrow \infty$ .
- ▶ Access to **limiting performances** and not only bounds!  $\Rightarrow$  **hyperparameter optimization, algorithm improvement.**
- ▶ **Complete intuitive change**  $\Rightarrow$  **opens way to renewed methods.**
- ▶ **Strong coincidence with real datasets**



## Take-away messages

- ▶ Asymptotic “**concentration effect**” for large  $n, p \Rightarrow$  **simplification in analyses and models.**
- ▶ Non-trivial **phase transition** phenomena (ability to detect, estimate) when  $p, n \rightarrow \infty$ .
- ▶ Access to **limiting performances** and not only bounds!  $\Rightarrow$  **hyperparameter optimization, algorithm improvement.**
- ▶ **Complete intuitive change**  $\Rightarrow$  **opens way to renewed methods.**
- ▶ **Strong coincidence with real datasets**  $\Rightarrow$  **easy link between theory and practice.**

- ▶ Neural nets: loss landscape, gradient descent dynamics and

- ▶ Neural nets: loss landscape, gradient descent dynamics and **deep learning!**

- ▶ Neural nets: loss landscape, gradient descent dynamics and **deep learning!**
- ▶ Generalized linear models

- ▶ Neural nets: loss landscape, gradient descent dynamics and **deep learning!**
- ▶ Generalized linear models
- ▶ More general problems from convex optimization (often of *implicit solution*)

- ▶ Neural nets: loss landscape, gradient descent dynamics and **deep learning!**
- ▶ Generalized linear models
- ▶ More general problems from convex optimization (often of *implicit solution*)
- ▶ More difficult: problem raised from *non-convex* optimization problems

- ▶ Neural nets: loss landscape, gradient descent dynamics and **deep learning!**
- ▶ Generalized linear models
- ▶ More general problems from convex optimization (often of *implicit solution*)
- ▶ More difficult: problem raised from *non-convex* optimization problems
- ▶ Transfer learning, active learning, generative networks (GAN)

- ▶ Neural nets: loss landscape, gradient descent dynamics and **deep learning!**
- ▶ Generalized linear models
- ▶ More general problems from convex optimization (often of *implicit solution*)
- ▶ More difficult: problem raised from *non-convex* optimization problems
- ▶ Transfer learning, active learning, generative networks (GAN)
- ▶ Robust statistics in machine learning



- ▶ Neural nets: loss landscape, gradient descent dynamics and **deep learning!**
- ▶ Generalized linear models
- ▶ More general problems from convex optimization (often of *implicit solution*)
- ▶ More difficult: problem raised from *non-convex* optimization problems
- ▶ Transfer learning, active learning, generative networks (GAN)
- ▶ Robust statistics in machine learning
- ▶ ...

Thank you.