Random Matrices in Machine Learning

Romain COUILLET

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

June 21, 2018

Outline

Outline

Outline

Context

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

Context

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

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

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

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

Context

Baseline scenario: $y_1,\ldots,y_n \in \mathbb{C}^p$ (or $\mathbb{R}^p$) i.i.d. with $E[y_1]=0$, $E[y_1y_1^*]=C_p$:

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

$$\hat{C}_p = \frac{1}{n}Y_pY_p^* = \frac{1}{n}\sum_{i=1}^{n} y_iy_i^*$$

  ($Y_p=[y_1,\ldots,y_n]\in\mathbb{C}^{p\times n}$).
- If $n\to\infty$, then, strong law of large numbers

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

  or equivalently, in spectral norm

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

Context

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

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

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

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

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

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

  or equivalently, in spectral norm

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

Random Matrix Regime

- No longer valid if $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

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

Context

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

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

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

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

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

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

  or equivalently, in spectral norm

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

Random Matrix Regime

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

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

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

The Marčenko–Pastur law

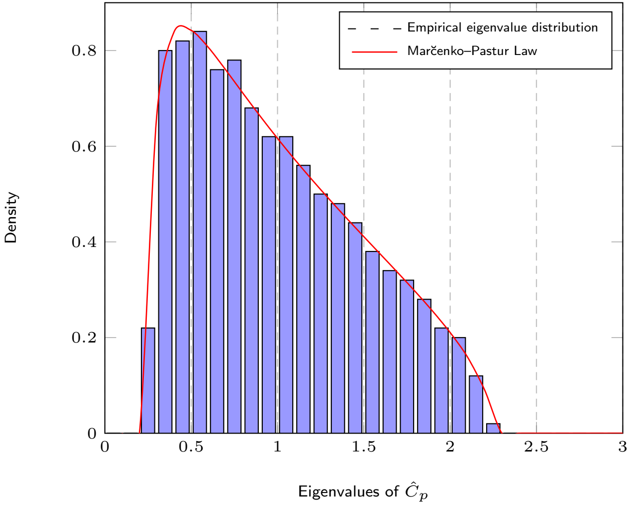


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $p = 500$, $n = 2000$, $C_p = I_p$.

Definition (Empirical Spectral Density)

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

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

Definition (Empirical Spectral Density)

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

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

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

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

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

weakly, where

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

Definition (Empirical Spectral Density)

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

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

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

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

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

weakly, where

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

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

The Marčenko–Pastur law

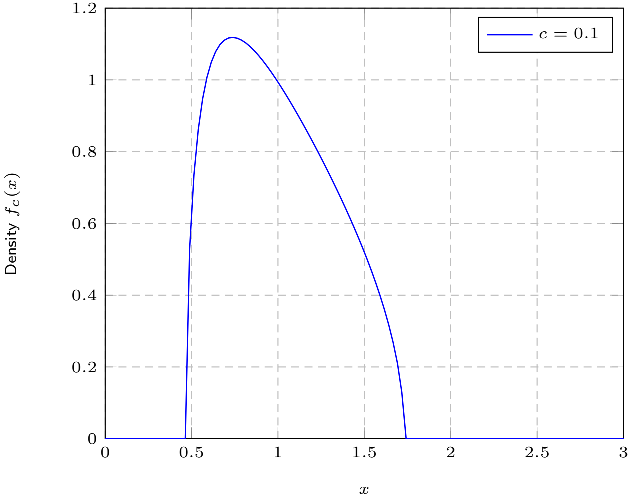


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

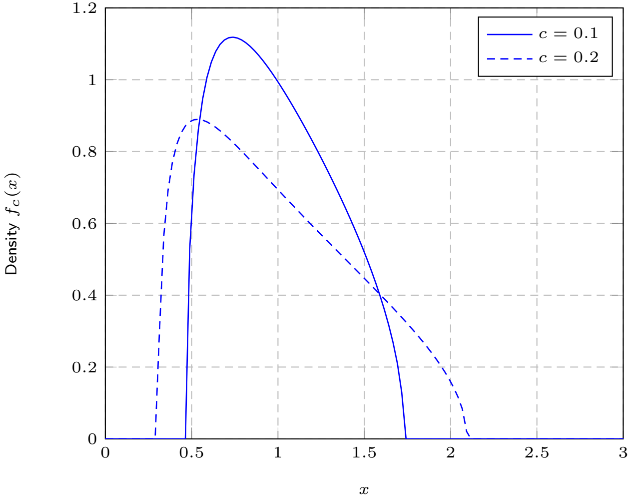


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

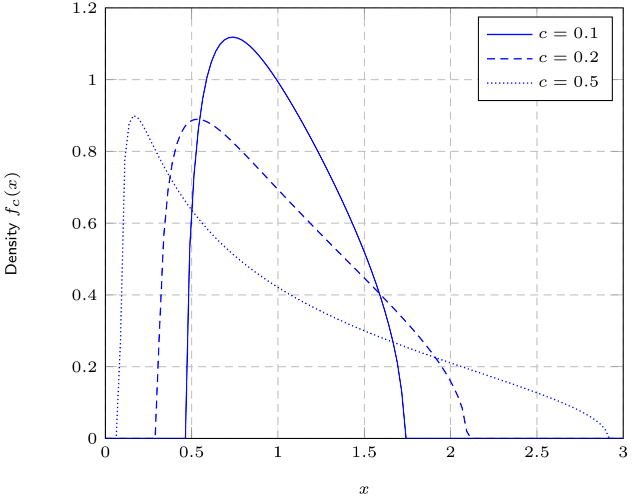


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

Outline

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

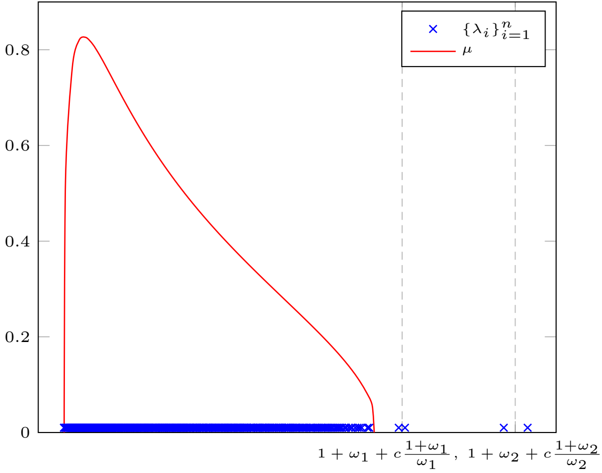


Figure: Eigenvalues of $\frac{1}{n}Y_pY_p^*$, $C_p = \text{diag}(\underbrace{1,\ldots,1}_{p-4}, 2, 2, 3, 3)$, $p = 500$, $n = 1500$.

Theorem (Eigenvalues [Baik,Silverstein'06])

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In particular,

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

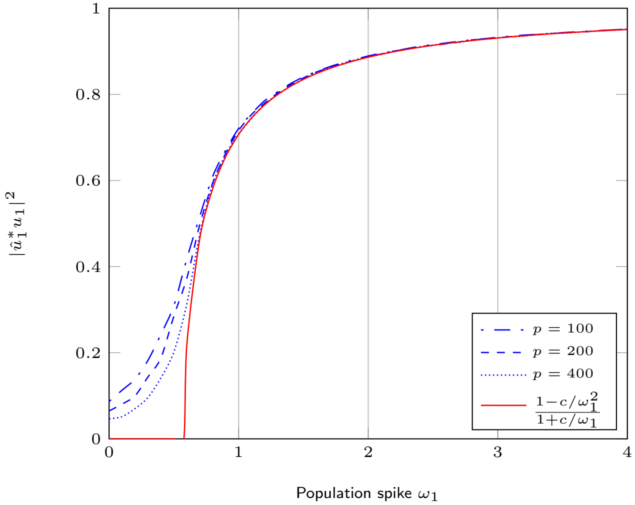


Figure: Simulated versus limiting $|\hat{u}_1^* u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^*$, $p/n = 1/3$, varying ω_1.

Other Spiked Models

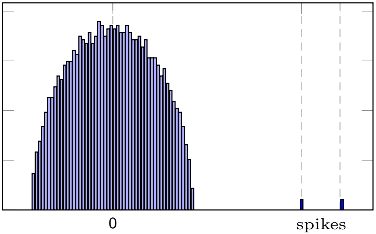
Similar results for multiple matrix models:

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

Outline

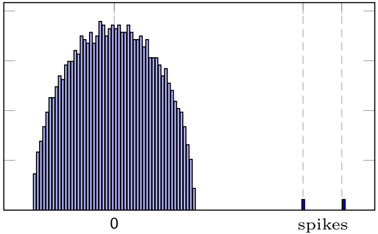
Outline

Context: Two-step classification of n objects based on similarity $A \in \mathbb{R}^{n \times n}$:

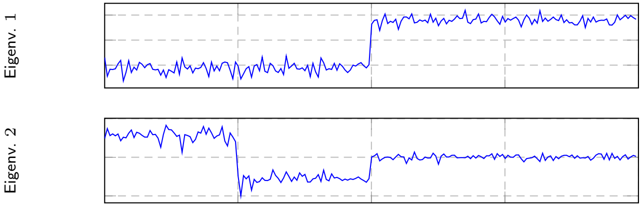


Reminder on Spectral Clustering Methods

Context: Two-step classification of n objects based on similarity $A \in \mathbb{R}^{n \times n}$:



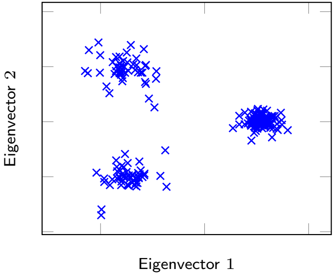
⇓ **Eigenvectors** ⇓
(in practice, shuffled)



Reminder on Spectral Clustering Methods

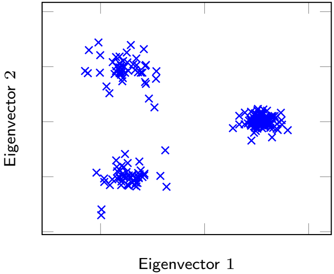
Reminder on Spectral Clustering Methods

⇓ ℓ**-dimensional representation** ⇓
(shuffling no longer matters)



Reminder on Spectral Clustering Methods

$\Downarrow$ **ℓ-dimensional representation** $\Downarrow$
(shuffling no longer matters)



$\Downarrow$
EM or k-means clustering.

Outline

Kernel Spectral Clustering

Problem Statement

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

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

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

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

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

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

Kernel Spectral Clustering

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

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

- Usually, $\kappa(x, y) = f(x^\mathsf{T} y)$ or $\kappa(x, y) = f(\|x - y\|^2)$
- Refinements:
  - instead of K, use $D - K$, $I_n - D^{-1}K$, $I_n - D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$, etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

Kernel Spectral Clustering

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

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

- Usually, $\kappa(x,y) = f(x^\mathsf{T} y)$ or $\kappa(x,y) = f(\|x-y\|^2)$
- Refinements:
  - instead of K, use $D - K$, $I_n - D^{-1}K$, $I_n - D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$, etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

Intuition (from small dimensions)

$$K = \begin{array}{cc} \left(\begin{array}{ccc} \kappa(x_i,x_j) \gg 1 & \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \ll 1 \\ \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \gg 1 & \kappa(x_i,x_j) \ll 1 \\ \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \gg 1 \end{array} \right) & \begin{array}{c} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \mathcal{C}_3 \end{array} \end{array}$$

- K essentially low rank with class structure in eigenvectors.

Kernel Spectral Clustering

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

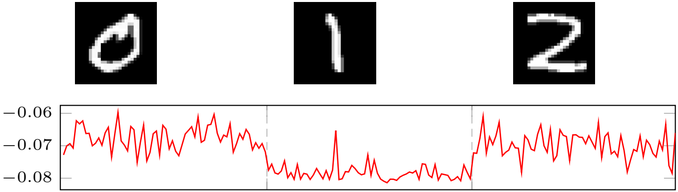
$$\boxed{K = \{\kappa(x_i, x_j)\}_{i,j=1}^n}$$

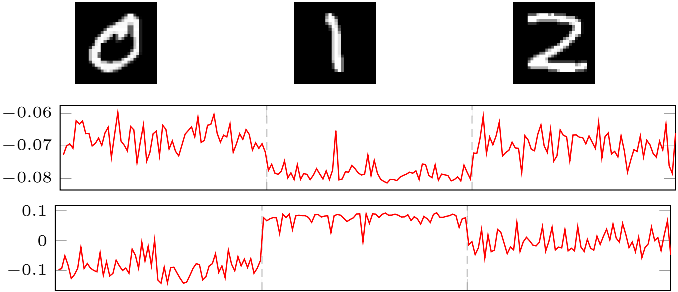
- Usually, $\kappa(x, y) = f(x^\mathsf{T} y)$ or $\kappa(x, y) = f(\|x - y\|^2)$
- Refinements:
  - instead of K, use $D - K$, $I_n - D^{-1}K$, $I_n - D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$, etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

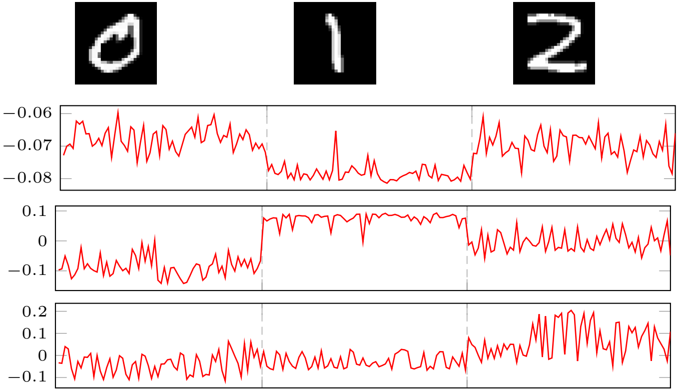
Intuition (from small dimensions)

$$K = \begin{array}{cc} \left(\begin{array}{ccc} \kappa(x_i,x_j) \gg 1 & \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \ll 1 \\ \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \gg 1 & \kappa(x_i,x_j) \ll 1 \\ \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \gg 1 \end{array} \right) & \begin{array}{c} \updownarrow \mathcal{C}_1 \\ \updownarrow \mathcal{C}_2 \\ \updownarrow \mathcal{C}_3 \end{array} \end{array}$$

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


−0.06
−0.07
−0.08


−0.06
−0.07
−0.08
0.1
0
−0.1


−0.06
−0.07
−0.08
0.1
0
−0.1
0.2
0.1
0
−0.1

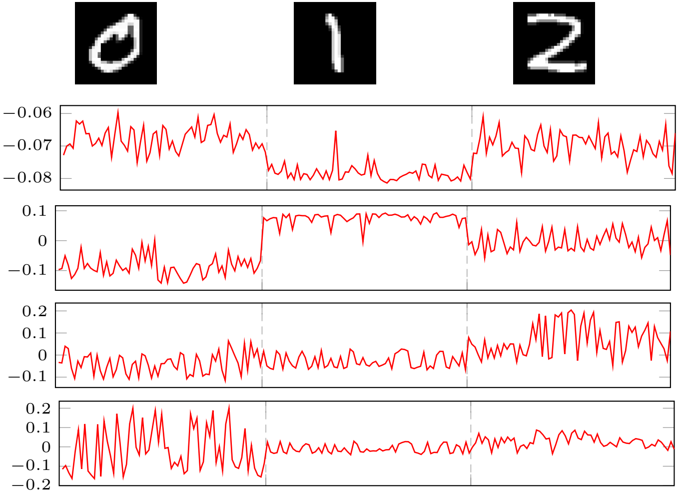


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data, RBF kernel ($f(t) = \exp(-t^2/2)$).

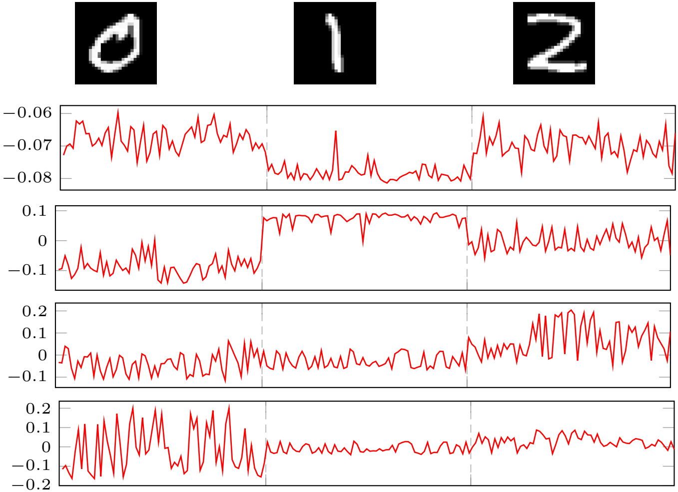


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data, RBF kernel ($f(t) = \exp(-t^2/2)$).

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

Model and Assumptions

Gaussian mixture model:

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

Model and Assumptions

Gaussian mixture model:

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

Assumption (Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,
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*

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

Gaussian mixture model:

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

Assumption (Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,
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*

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

For 2 classes, this is

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

Gaussian mixture model:

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

Assumption (Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0,1)$,
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*

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

For 2 classes, this is

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

Remark: [Neyman–Pearson optimality]

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

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 ($f(\frac{1}{p} x_i^\mathsf{T} x_j)$ 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 ($f(\frac{1}{p} x_i^\mathsf{T} x_j)$ simpler).

- We study the normalized Laplacian:

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

  with $d = K1_n$, $D = \mathrm{diag}(d)$.
  (more stable both theoretically and in practice)

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

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

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

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

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

where $\tau=\frac{1}{p}\operatorname{tr}C^\circ$.
$\Rightarrow$ Suggests that (up to diagonal) $K\simeq f(\tau)1_n1_n^\mathsf{T}$!

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

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

  where $\tau=\frac{1}{p}\operatorname{tr}C^\circ$.
  $\Rightarrow$ Suggests that (up to diagonal) $K\simeq f(\tau)1_n1_n^\mathsf{T}$!

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

$$K=\underbrace{f(\tau)1_n1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)}+\underbrace{\sqrt{n}K_1}_{\text{low rank, }O_{\|\cdot\|}(\sqrt{n})}+\underbrace{K_2}_{\text{informative terms, }O_{\|\cdot\|}(1)}$$

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

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

  where $\tau=\frac{1}{p}\operatorname{tr}C^\circ$.
  $\Rightarrow$ Suggests that (up to diagonal) $K\simeq f(\tau)1_n1_n^\mathsf{T}$!

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

$$K=\underbrace{f(\tau)1_n1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)}+\underbrace{\sqrt{n}K_1}_{\text{low rank, }O_{\|\cdot\|}(\sqrt{n})}+\underbrace{K_2}_{\text{informative terms, }O_{\|\cdot\|}(1)}$$

Clearly not the (small dimension) expected behavior.

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

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

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

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

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

Theorem (Random Matrix Equivalent [Couillet, Benaych'2015])
As $n, p \to \infty$, $\left\| L - \hat{L} \right\| \xrightarrow{\text{a.s.}} 0$, *where*

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

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

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

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

$$B = M^{\mathsf{T}}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) tt^{\mathsf{T}} - \frac{f''(\tau)}{f'(\tau)} T + *.$$

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

Theorem (Random Matrix Equivalent **[Couillet, Benaych'2015]**)
As $n,p\to\infty$, $\left\|L-\hat{L}\right\| \xrightarrow{\text{a.s.}} 0$, *where*

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

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

et $W = [w_1,\ldots,w_n] \in \mathbb{R}^{p\times n}$ *(*$x_i = \mu_a + w_i$*)*, $P = I_n - \frac{1}{n}1_n1_n^{\mathsf{T}}$,

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

$$B = M^{\mathsf{T}}M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)}\right)tt^{\mathsf{T}} - \frac{f''(\tau)}{f'(\tau)}T + *.$$

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

Fundamental conclusions:

Theorem (Random Matrix Equivalent [Couillet, Benaych'2015])
As $n, p \to \infty$, $\left\| L - \hat{L} \right\| \xrightarrow{\text{a.s.}} 0$, *where*

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

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

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

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

$$B = M^\mathsf{T} M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) tt^\mathsf{T} - \frac{f''(\tau)}{f'(\tau)} T + *.$$

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

Fundamental conclusions:

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

Theorem (Random Matrix Equivalent [Couillet, Benaych'2015])
As $n, p \to \infty$, $\left\| L - \hat{L} \right\| \xrightarrow{\text{a.s.}} 0$, *where*

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

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

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

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

$$B = M^\mathsf{T} M + \left(\frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) tt^\mathsf{T} - \frac{f''(\tau)}{f'(\tau)} T + *.$$

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

Fundamental conclusions:

- asymptotic **kernel impact** only through $f'(\tau)$ and $f''(\tau)$, **that's all!**
- spectral clustering **reads** $M^\mathsf{T} M$, tt^T and T, **that's all!**

Isolated eigenvalues: Gaussian inputs

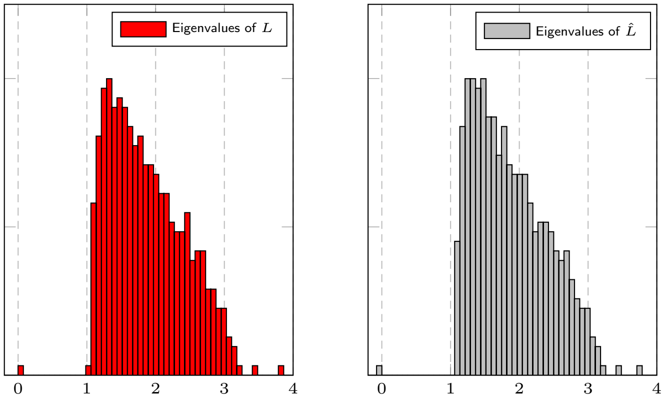


Figure: Eigenvalues of L and $\hat{L}$, $k = 3$, $p = 2048$, $n = 512$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $[\mu_a]_j = 4\boldsymbol{\delta}_{aj}$, $C_a = (1 + 2(a-1)/\sqrt{p})I_p$, $f(x) = \exp(-x/2)$.

Theoretical Findings versus MNIST

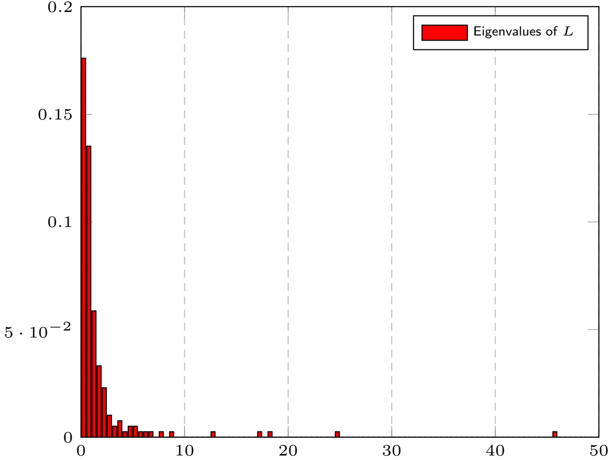


Figure: Eigenvalues of L (red) and (equivalent Gaussian model) $\hat{L}$ (white), MNIST data, $p = 784$, $n = 192$.

Theoretical Findings versus MNIST

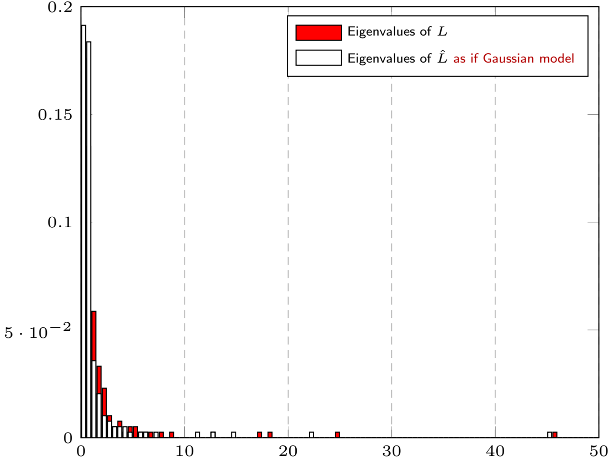


Figure: Eigenvalues of L (red) and (equivalent Gaussian model) $\hat{L}$ (white), MNIST data, $p = 784$, $n = 192$.

Theoretical Findings versus MNIST

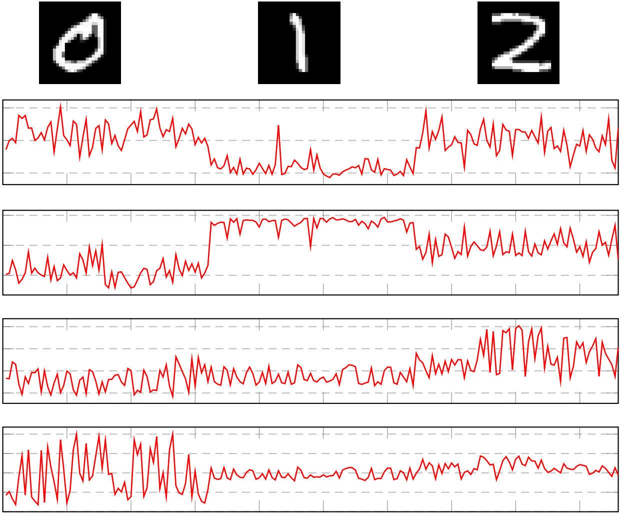


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theoretical Findings versus MNIST

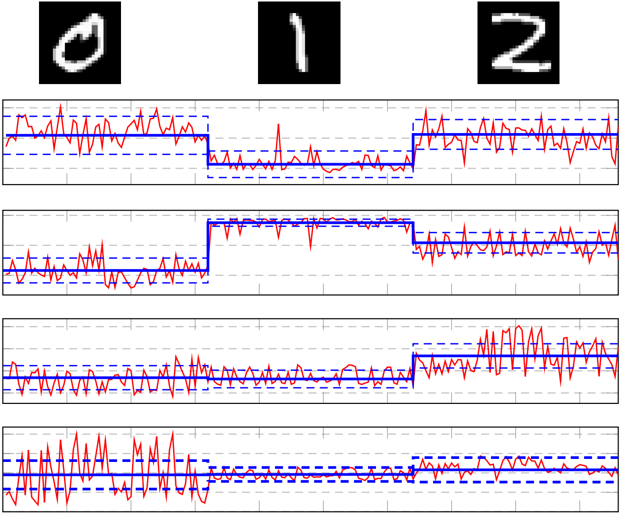


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theoretical Findings versus MNIST

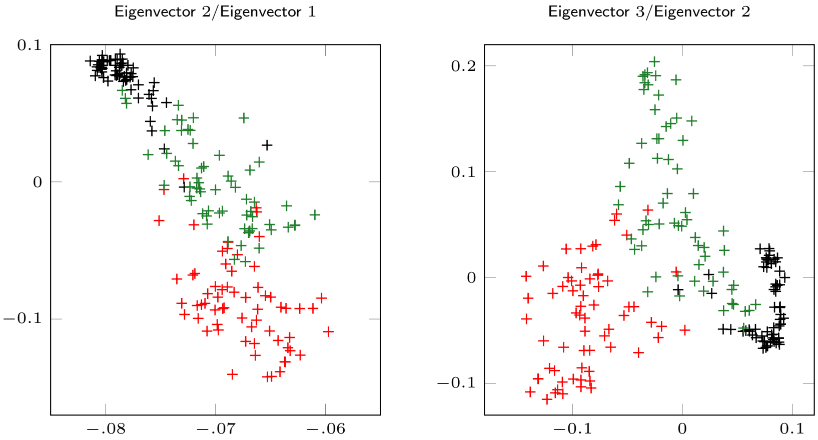


Figure: 2D representation of eigenvectors of L, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

Theoretical Findings versus MNIST

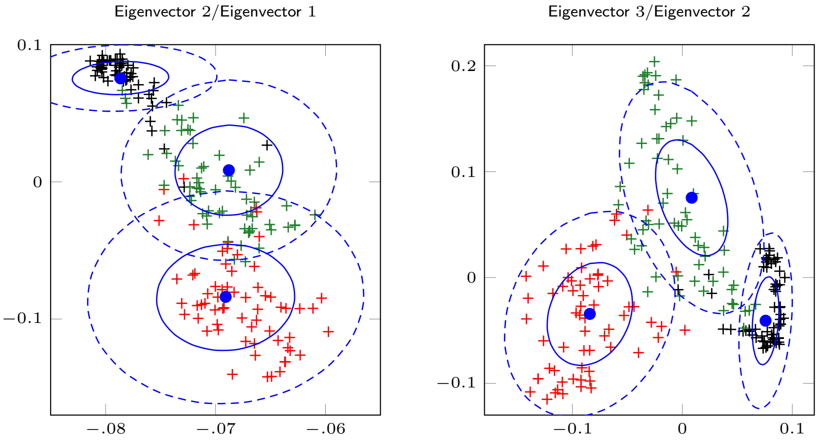


Figure: 2D representation of eigenvectors of L, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

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

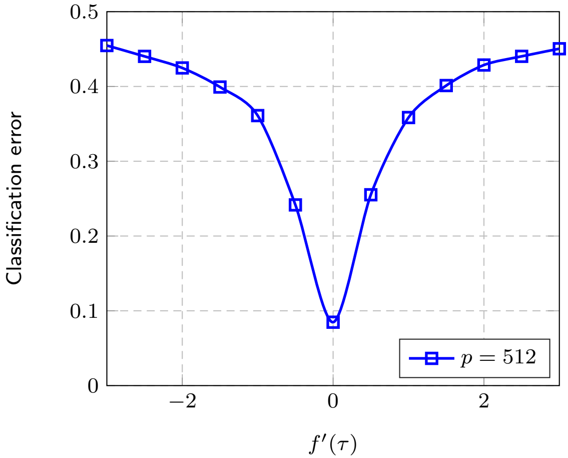


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

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

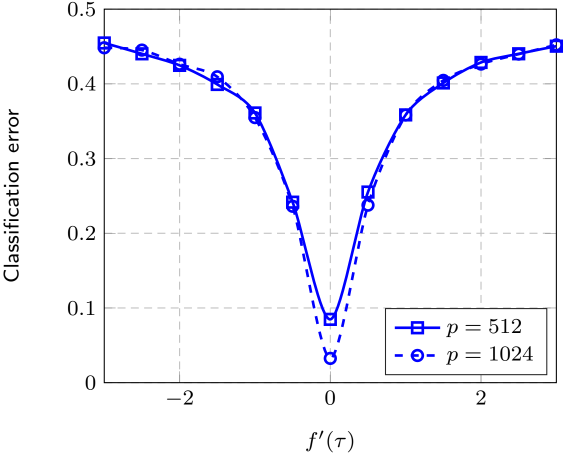


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

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

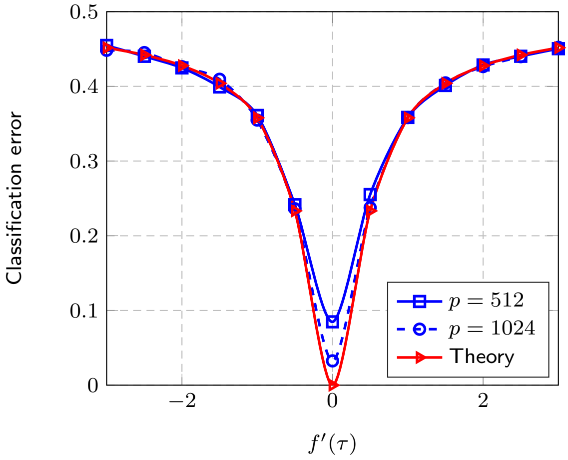


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

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

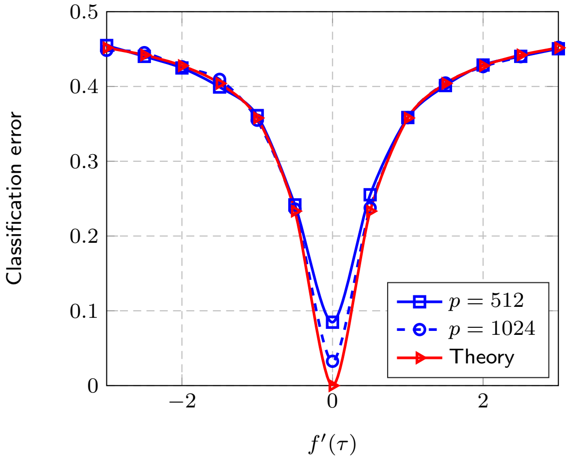


Figure: Polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

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

Outline

Problem: Cluster large data $x_1, \ldots, x_n \in \mathbb{R}^p$ based on "spanned subspaces".

Position of the Problem

Problem: Cluster large data $x_1, \ldots, x_n \in \mathbb{R}^p$ based on "spanned subspaces".

Method:

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

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

Position of the Problem

Problem: Cluster large data $x_1, \ldots, x_n \in \mathbb{R}^p$ based on "spanned subspaces".

Method:

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

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

- Performance of $L = nD^{-\frac{1}{2}} \left(K - \frac{1_n 1_n^\mathsf{T}}{1_n^\mathsf{T} D 1_n} \right) D^{-\frac{1}{2}}$, 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 $n, p \to \infty$.
(alternatively, we can ask $\frac{1}{p} \mathrm{tr}\, C_i = 1$ *for all* $1 \leq i \leq k$*)*

Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ 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 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ 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 $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

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

Assumption 1 [Classes]. Vectors $x_1,\ldots,x_n\in\mathbb{R}^p$ 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 $n\to\infty$, for each $a\in\{1,\ldots,k\}$,

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

Theorem (Corollary of Previous Section)

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

$$L=nD^{-\frac{1}{2}}\left(K-\frac{1_n1_n^\mathsf{T}}{1_n^\mathsf{T}D1_n}\right)D^{-\frac{1}{2}},\ \textit{with}\ K=\left\{f\left(\|\bar{x}_i-\bar{x}_j\|^2\right)\right\}_{i,j=1}^n\ \ (\bar{x}=x/\|x\|)$$

exhibits phase transition phenomenon

Assumption 1 [Classes]. Vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ 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 $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(p)$, 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,

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

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

$$T = \left\{ \frac{1}{p} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k$$

has sufficiently large eigenvalues (here $M = 0$, $t = 0$).

The case $f'(2) = 0$

Assumption 2b [Growth Rates]. As $n \to \infty$, for each $a \in \{1, \ldots, k\}$,

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

The case $f'(2) = 0$

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

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$, 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 $\|E\| \to 0$,* **detectability iif** $\frac{1}{p} tr (C_1 - C_2)^2 \geq O(p^{-\frac{1}{2}})$.

The case $f'(2) = 0$

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

1. $\frac{n}{p} \to c_0 \in (0, \infty)$
2. $\frac{n_a}{n} \to c_a \in (0, \infty)$
3. $\frac{1}{p} \operatorname{tr} C_a = 1$ and $\operatorname{tr} C_a^\circ C_b^\circ = O(\sqrt{p})$, 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* $\|E\| \to 0$, **detectability iif** $\frac{1}{p} tr (C_1 - C_2)^2 \geq O(p^{-\frac{1}{2}})$.

Theorem (Random Equivalent for $f'(2) = 0$)

Let f be smooth with $f'(2) = 0$ and

$$\mathcal{L} \equiv \sqrt{p} \frac{f(2)}{2f''(2)} \left[L - \frac{f(0) - f(2)}{f(2)} P \right], \quad P = I_n - \frac{1}{n} 1_n 1_n^\mathsf{T}.$$

Then, under Assumptions 2b,

$$\mathcal{L} = P \Phi P + \left\{ \frac{1}{\sqrt{p}} tr (C_a^\circ C_b^\circ) \frac{1_{n_a} 1_{n_b}^\mathsf{T}}{p} \right\}_{a,b=1}^k + o_{\|\cdot\|}(1)$$

where $\Phi_{ij} = \boldsymbol{\delta}_{i \neq j} \sqrt{p} \left[(x_i^\mathsf{T} x_j)^2 - E[(x_i^\mathsf{T} x_j)^2] \right]$.

The case $f'(2) = 0$

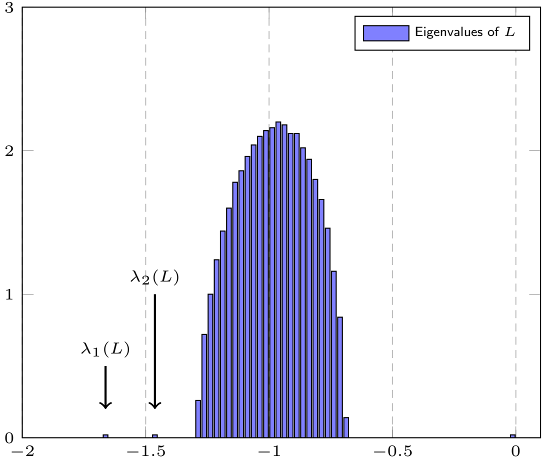
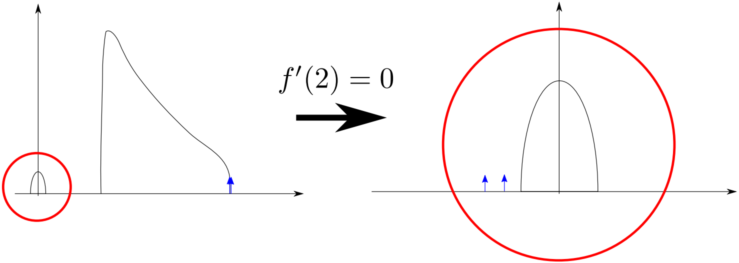


Figure: Eigenvalues of L, $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^\mathsf{T}$, $W_i \in \mathbb{R}^{p \times (p/8)}$ of i.i.d. $\mathcal{N}(0,1)$ entries, $f(t) = \exp(-(t-2)^2)$.

$\Rightarrow$ **No longer a Marcenko–Pastur like bulk, but rather a semi-circle bulk!**

The case $f'(2) = 0$



The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ

The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ
- study the isolated eigenvalues of $\mathcal{L}$ (and the phase transition)

The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ
- study the isolated eigenvalues of $\mathcal{L}$ (and the phase transition)
- retrieve information from the eigenvectors.

The case $f'(2) = 0$

Roadmap. We now need to:

- study the spectrum of Φ
- study the isolated eigenvalues of $\mathcal{L}$ (and the phase transition)
- retrieve information from the eigenvectors.

Theorem (Semi-circle law for Φ)

Let $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}_{\lambda_i(\mathcal{L})}$. Then, under Assumption 2b,

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

with μ the semi-circle distribution

$$\mu(dt) = \frac{1}{2\pi c_0 \omega^2}\sqrt{(4c_0\omega^2 - t^2)^+}dt, \quad \omega = \lim_{p\to\infty} \sqrt{2}\frac{1}{p}\mathit{tr}(C^\circ)^2.$$

The case $f'(2) = 0$

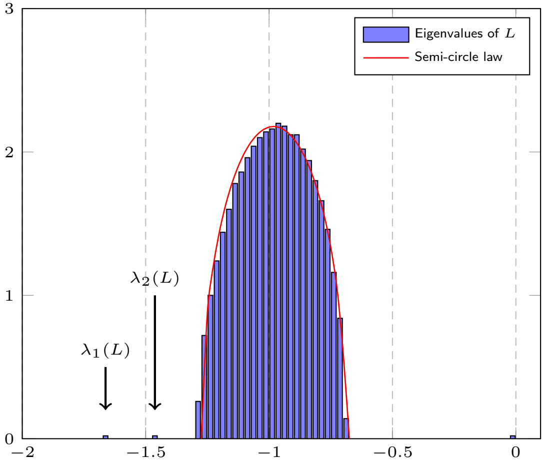


Figure: Eigenvalues of L, $p = 1000$, $n = 2000$, $k = 3$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^{\mathsf{T}}$, $W_i \in \mathbb{R}^{p\times(p/8)}$ of i.i.d. $\mathcal{N}(0,1)$ entries, $f(t) = \exp(-(t-2)^2)$.

The case $f'(2) = 0$

Denote now

$$\mathcal{T} \equiv \lim_{p\to\infty} \left\{ \frac{\sqrt{c_a c_b}}{\sqrt{p}} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^{k}.$$

Denote now

$$\mathcal{T} \equiv \lim_{p\to\infty} \left\{ \frac{\sqrt{c_a c_b}}{\sqrt{p}} \operatorname{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^{k}.$$

Theorem (Isolated Eigenvalues)

Let $\nu_1 \geq \ldots \geq \nu_k$ eigenvalues of $\mathcal{T}$. Then, if $\sqrt{c_0}|\nu_i| > \omega$, $\mathcal{L}$ has an isolated eigenvalue λ_i satisfying

$$\lambda_i \xrightarrow{\text{a.s.}} \rho_i \equiv c_0 \nu_i + \frac{\omega^2}{\nu_i}.$$

Theorem (Isolated Eigenvectors)

For each isolated eigenpair (λ_i, u_i) *of* $\mathcal{L}$ *corresponding to* (ν_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 $j_a = [0_{n_1}^{\mathsf{T}}, \ldots, 1_{n_a}^{\mathsf{T}}, \ldots, 0_{n_k}^{\mathsf{T}}]^{\mathsf{T}}$, $(w_i^a)^{\mathsf{T}} j_a = 0$, $\operatorname{supp}(w_i^a) = \operatorname{supp}(j_a)$, $\|w_i^a\| = 1$.
Then, under Assumptions 1–2b,

$$\alpha_i^a \alpha_i^b \xrightarrow{\text{a.s.}} \left(1 - \frac{1}{c_0} \frac{\omega^2}{\nu_i^2}\right) [v_i v_i^{\mathsf{T}}]_{ab}$$

$$(\sigma_i^a)^2 \xrightarrow{\text{a.s.}} \frac{c_a}{c_0} \frac{\omega^2}{\nu_i^2}$$

and the fluctuations of u_i, u_j*,* $i \neq j$*, are asymptotically uncorrelated.*

The case $f'(2) = 0$

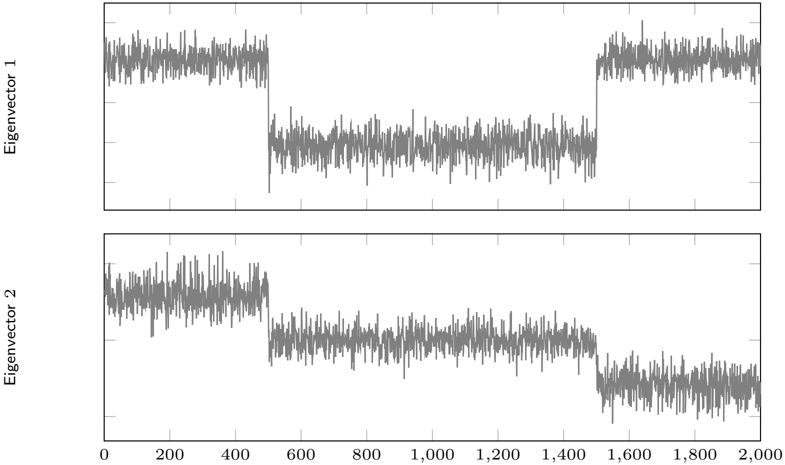


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

The case $f'(2) = 0$

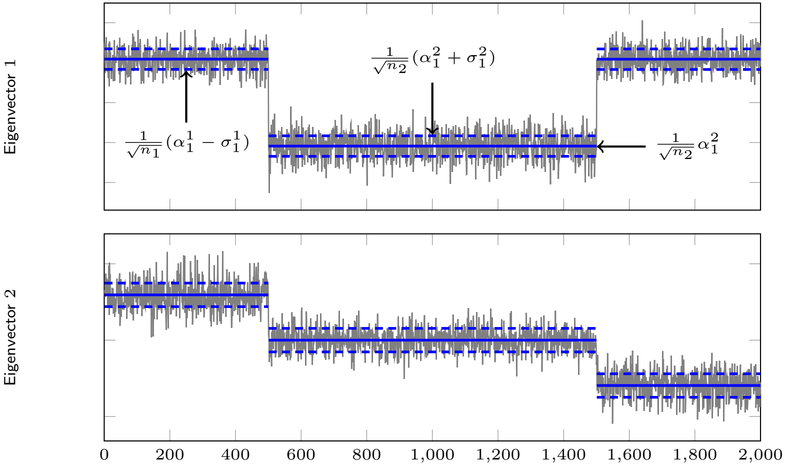


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

The case $f'(2) = 0$

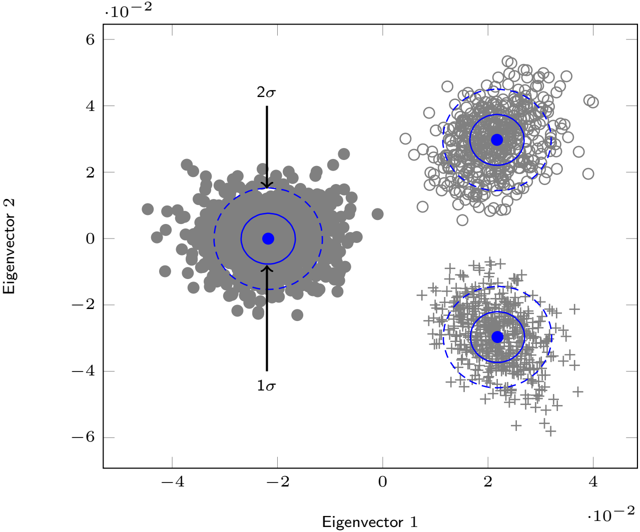
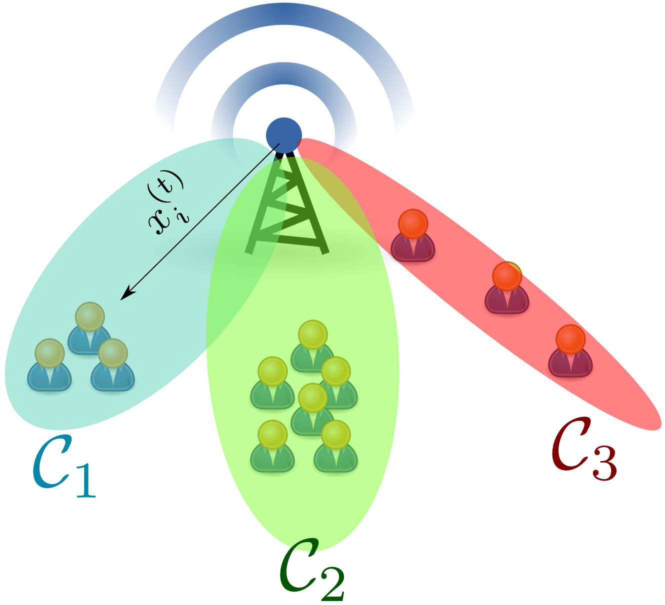


Figure: Leading two eigenvectors of $\mathcal{L}$ (or equivalently of L) versus deterministic approximations of $\alpha_i^a \pm \sigma_i^a$.

Application to Massive MIMO UE Clustering



Massive MIMO UE Clustering

Setting. Massive MIMO cell with

- p antenna elements
- n users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE's belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^\mathsf{T}] = C_a \equiv C(\Theta_a)$.

Massive MIMO UE Clustering

Setting. Massive MIMO cell with

- p antenna elements
- n users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE's belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^{\mathsf{T}}] = C_a \equiv C(\Theta_a)$.
- T independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE i.

Massive MIMO UE Clustering

Setting. Massive MIMO cell with

- p antenna elements
- n users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE's belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^{\mathsf{T}}] = C_a \equiv C(\Theta_a)$.
- T independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE i.

Objective. Clustering users in same solid angle groups *(for scheduling reasons, to avoid pilot contamination)*.

Massive MIMO UE Clustering

Setting. Massive MIMO cell with

- p antenna elements
- n users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE's belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^\mathsf{T}] = C_a \equiv C(\Theta_a)$.
- T independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE i.

Objective. Clustering users in same solid angle groups *(for scheduling reasons, to avoid pilot contamination)*.

Algorithm.

1. Build kernel matrix K, then $\mathcal{L}$, based on nT vectors $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if nT values to cluster).

Massive MIMO UE Clustering

Setting. Massive MIMO cell with

- p antenna elements
- n users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE's belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^{\mathsf{T}}] = C_a \equiv C(\Theta_a)$.
- T independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE i.

Objective. Clustering users in same solid angle groups *(for scheduling reasons, to avoid pilot contamination)*.

Algorithm.

1. Build kernel matrix K, then $\mathcal{L}$, based on nT vectors $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if nT values to cluster).
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$

Massive MIMO UE Clustering

Setting. Massive MIMO cell with

- p antenna elements
- n users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE's belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^\mathsf{T}] = C_a \equiv C(\Theta_a)$.
- T independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE i.

Objective. Clustering users in same solid angle groups *(for scheduling reasons, to avoid pilot contamination)*.

Algorithm.

1. Build kernel matrix K, then $\mathcal{L}$, based on nT vectors $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if nT values to cluster).
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$
3. For each i, create $\tilde{u}_i = \frac{1}{T}(I_n \otimes 1_T^\mathsf{T})u_i$, i.e., average eigenvectors along time.

Massive MIMO UE Clustering

Setting. Massive MIMO cell with

- p antenna elements
- n users equipments (UE) with channels $x_1, \ldots, x_n \in \mathbb{R}^p$
- UE's belong to solid angle groups, i.e., $E[x_i] = 0$, $E[x_i x_i^{\mathsf{T}}] = C_a \equiv C(\Theta_a)$.
- T independent channel observations $x_i^{(1)}, \ldots, x_i^{(T)}$ for UE i.

Objective. Clustering users in same solid angle groups *(for scheduling reasons, to avoid pilot contamination)*.

Algorithm.

1. Build kernel matrix K, then $\mathcal{L}$, based on nT vectors $x_1^{(1)}, \ldots, x_n^{(T)}$ (as if nT values to cluster).
2. Extract dominant isolated eigenvectors $u_1, \ldots, u_\kappa$
3. For each i, create $\tilde{u}_i = \frac{1}{T}(I_n \otimes 1_T^{\mathsf{T}})u_i$, i.e., average eigenvectors along time.
4. Perform k-class clustering on vectors $\tilde{u}_1, \ldots, \tilde{u}_\kappa$.

Massive MIMO UE Clustering

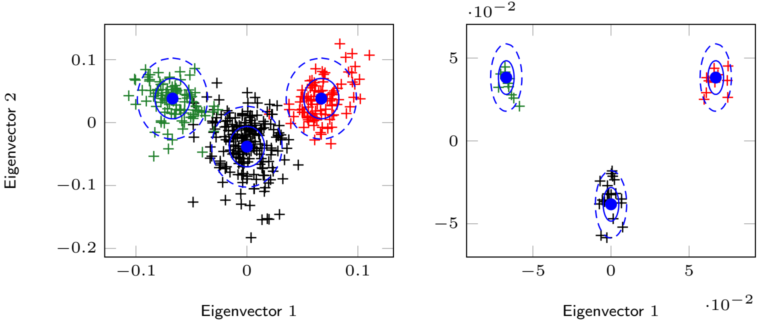


Figure: Leading two eigenvectors before (left figure) and after (right figure) T-averaging. Setting: $p = 400$, $n = 40$, $T = 10$, $k = 3$, $c_1 = c_3 = 1/4$, $c_2 = 1/2$, angular spread model with angles $-\pi/30 \pm \pi/20$, $0 \pm \pi/20$, and $\pi/30 \pm \pi/20$. Kernel function $f(t) = \exp(-(t-2)^2)$.

Massive MIMO UE Clustering

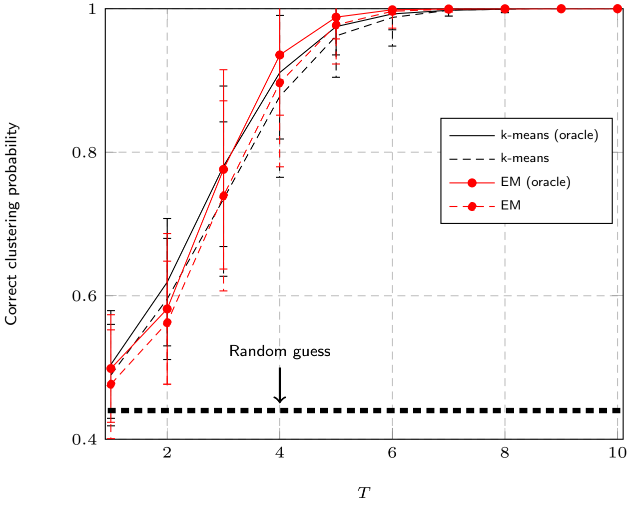


Figure: Overlap for different T, using the k-means or EM starting from actual centroid solutions (oracle) or randomly.

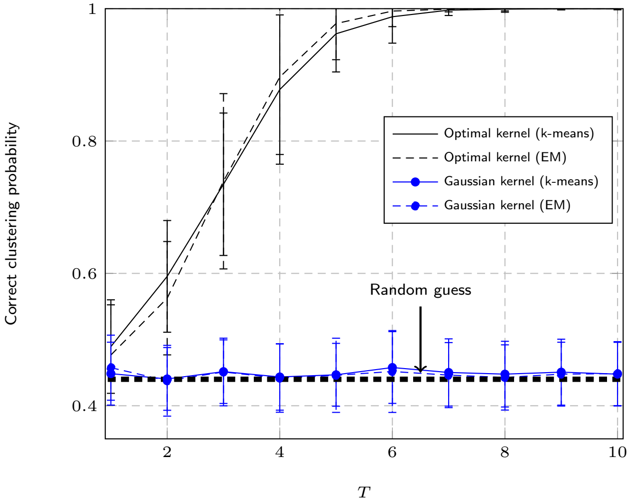


Figure: Overlap for optimal kernel $f(t)$ (here $f(t) = \exp(-(t-2)^2)$) and Gaussian kernel $f(t) = \exp(-t^2)$, for different T, using the k-means or EM.

Outline

Conclusion of previous analyses:

- kernel $f(\frac{1}{p}\|x_i - x_j\|^2)$ with $f'(\tau) \neq 0$:
  - optimal in $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\operatorname{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(1)$

$\longrightarrow$ **Model type:** Marčenko–Pastur + spikes.

Conclusion of previous analyses:

- kernel $f(\frac{1}{p}\|x_i - x_j\|^2)$ with $f'(\tau) \neq 0$:
  - optimal in $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\operatorname{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(1)$

$\longrightarrow$ **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**)
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(p^{-\frac{1}{2}})$

$\longrightarrow$ **Model type:** smaller order semi-circle law + spikes.

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 $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\operatorname{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(1)$

  $\longrightarrow$ **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)**
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(p^{-\frac{1}{2}})$

  $\longrightarrow$ **Model type:** smaller order semi-circle law + spikes.

Jointly optimal solution:

- evenly weighing Marčenko–Pastur and semi-circle laws

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 $\|\mu_a^\circ\| = O(1)$, $\frac{1}{p}\operatorname{tr} C_a^\circ = O(p^{-\frac{1}{2}})$
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(1)$

  $\longrightarrow$ **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**)
  - suboptimal in $\frac{1}{p}\operatorname{tr} C_a^\circ C_b^\circ = O(p^{-\frac{1}{2}})$

  $\longrightarrow$ **Model type:** smaller order semi-circle law + spikes.

Jointly optimal solution:

- evenly weighing Marčenko–Pastur and semi-circle laws
- the "α-β" kernel:

$$f'(\tau) = \frac{\alpha}{\sqrt{p}}, \quad \frac{1}{2}f''(\tau) = \beta.$$

New assumption setting

- We consider now a **fully optimal growth rate setting**

Assumption (Optimal Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$, $\frac{n_a}{n} \to c_a \in (0, 1)$,
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*

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

New assumption setting

- We consider now a **fully optimal growth rate setting**

Assumption (Optimal Growth Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0,\infty)$, $\frac{n_a}{n} \to c_a \in (0,1)$,
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*

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

Kernel:

- For technical simplicity, we consider

$$\boxed{\tilde{K} = PKP = P\left\{ f\left(\frac{1}{p}(x^\circ)^\mathsf{T}(x_j^\circ)\right)\right\}_{i,j=1}^n P} \quad P = I_n - \frac{1}{n}1_n 1_n^\mathsf{T}.$$

i.e., τ replaced by 0.

Theorem

As $n \to \infty$,

$$\left\| \sqrt{p} \left(PKP + \left(f(0) + \tau f'(0) \right) P \right) - \hat{\mathcal{K}} \right\| \xrightarrow{\text{a.s.}} 0$$

with, for $\alpha = \sqrt{p} f'(0) = O(1)$ *and* $\beta = \frac{1}{2} f''(0) = O(1)$,

$$\hat{\mathcal{K}} = \alpha P W^T W P + \beta P \Phi P + U A U^T$$

$$A = \begin{bmatrix} \alpha M^T M + \beta T & \alpha I_k \\ \alpha I_k & 0 \end{bmatrix}$$

$$U = \left[\frac{J}{\sqrt{p}}, P W^T M \right]$$

$$\frac{\Phi}{\sqrt{p}} = \left\{ ((\omega_i^\circ)^T \omega_j^\circ)^2 \boldsymbol{\delta}_{i \neq j} \right\}_{i,j=1}^{n} - \left\{ \frac{\mathrm{tr}\,(C_a C_b)}{p^2} 1_{n_a} 1_{n_b}^T \right\}_{a,b=1}^{k}.$$

Theorem

As $n \to \infty$,

$$\left\| \sqrt{p}\left(PKP + \left(f(0) + \tau f'(0)\right) P\right) - \hat{\mathcal{K}} \right\| \xrightarrow{\text{a.s.}} 0$$

with, for $\alpha = \sqrt{p} f'(0) = O(1)$ *and* $\beta = \frac{1}{2} f''(0) = O(1)$,

$$\hat{\mathcal{K}} = \alpha P W^T W P + \beta P \Phi P + U A U^T$$

$$A = \begin{bmatrix} \alpha M^T M + \beta T & \alpha I_k \\ \alpha I_k & 0 \end{bmatrix}$$

$$U = \left[\frac{J}{\sqrt{p}}, P W^T M \right]$$

$$\frac{\Phi}{\sqrt{p}} = \left\{ ((\omega_i^\circ)^T \omega_j^\circ)^2 \boldsymbol{\delta}_{i \neq j} \right\}_{i,j=1}^{n} - \left\{ \frac{tr\,(C_a C_b)}{p^2} 1_{n_a} 1_{n_b}^T \right\}_{a,b=1}^{k}.$$

Role of α**,** β:

- Weighs Marčenko–Pastur versus semi-circle parts.

Theorem (Eigenvalues Bulk)

As $p \to \infty$,

$$\nu_n \triangleq \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\delta}_{\lambda_i(\hat{K})} \xrightarrow{\text{a.s.}} \nu$$

with ν having Stieltjes transform $m(z)$ solution of

$$\frac{1}{m(z)} = -z + \frac{\alpha}{p} \operatorname{tr} C^\circ \left(I_k + \frac{\alpha m(z)}{c_0} C^\circ \right)^{-1} - \frac{2\beta^2}{c_0} \omega^2 m(z)$$

where $\omega = \lim_{p \to \infty} \frac{1}{p} \operatorname{tr} (C^\circ)^2$.

Limiting eigenvalue distribution

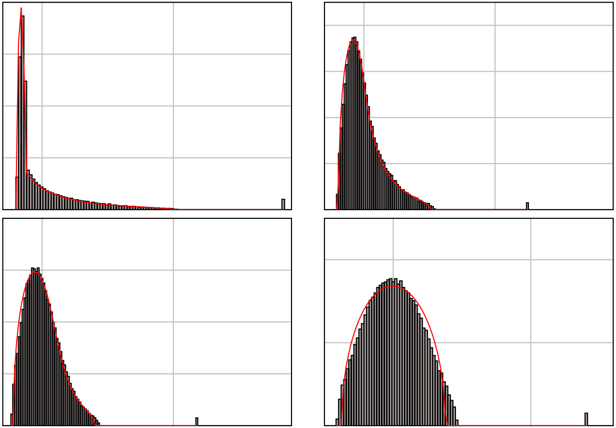


Figure: Eigenvalues of K (up to recentering) versus limiting law, $p = 2048$, $n = 4096$, $k = 2$, $n_1 = n_2$, $\boldsymbol{\mu}_i = 3\boldsymbol{\delta}_i$, $f(x) = \frac{1}{2}\beta\left(x + \frac{1}{\sqrt{p}}\frac{\alpha}{\beta}\right)^2$. **(Top left)**: $\alpha = 8, \beta = 1$, **(Top right)**: $\alpha = 4, \beta = 3$, **(Bottom left)**: $\alpha = 3, \beta = 4$, **(Bottom right)**: $\alpha = 1, \beta = 8$.

Asymptotic performances: MNIST

- MNIST is "means-dominant" but not that much!

Datasets	$\|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$	$\frac{1}{\sqrt{p}} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$
MNIST (digits $1, 7$)	612.7	71.1	2.5
MNIST (digits $3, 6$)	441.3	39.9	1.4
MNIST (digits $3, 8$)	212.3	23.5	0.8

Asymptotic performances: MNIST

- MNIST is "means-dominant" but not that much!

Datasets	$\|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$	$\frac{1}{\sqrt{p}} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$
MNIST (digits 1, 7)	612.7	71.1	2.5
MNIST (digits 3, 6)	441.3	39.9	1.4
MNIST (digits 3, 8)	212.3	23.5	0.8

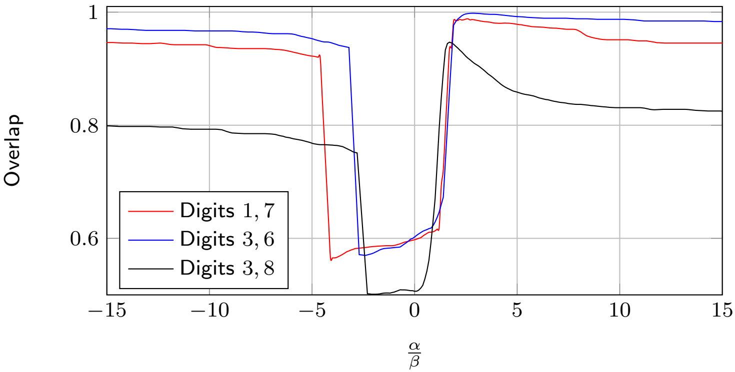


Figure: Spectral clustering of the MNIST database for varying $\frac{\alpha}{\beta}$.

Asymptotic performances: EEG data

- EEG data are "variance-dominant"

Datasets	$\|\boldsymbol{\mu}_1^\circ - \boldsymbol{\mu}_2^\circ\|^2$	$\frac{1}{\sqrt{p}} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \operatorname{tr} (\mathbf{C}_1 - \mathbf{C}_2)^2$
EEG (sets A, E)	2.4	10.9	1.1

- EEG data are "variance-dominant"

Datasets	$\|\mu_1^\circ - \mu_2^\circ\|^2$	$\frac{1}{\sqrt{p}} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$	$\frac{1}{p} \operatorname{tr}(\mathbf{C}_1 - \mathbf{C}_2)^2$
EEG (sets A, E)	2.4	10.9	1.1

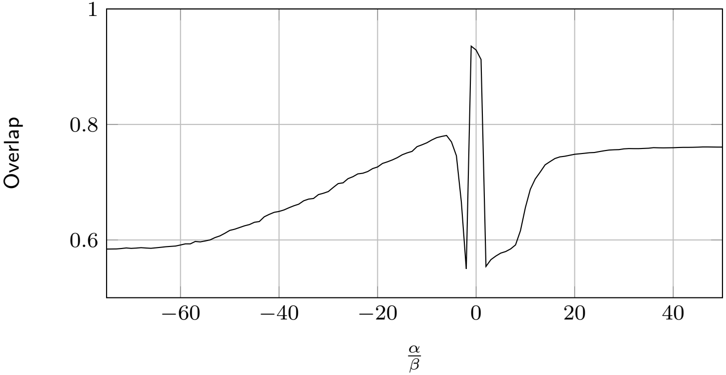


Figure: Spectral clustering of the EEG database for varying $\frac{\alpha}{\beta}$.

Outline

Problem Statement

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in k classes, with n_l labelled and n_u unlabelled data.

Problem Statement

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ 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 $F_{ia} = \boldsymbol{\delta}_{\{x_i \in \mathcal{C}_a\}}$, for all labelled x_i.

Problem Statement

Context: Similar to clustering:

- Classify $x_1,\ldots,x_n\in\mathbb{R}^p$ 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 $F_{ia}=\boldsymbol{\delta}_{\{x_i\in\mathcal{C}_a\}}$, 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

$$K=\begin{bmatrix}K_{(l,l)} & K_{(l,u)}\\ K_{(u,l)} & K_{(u,u)}\end{bmatrix}$$
$$D=\begin{bmatrix}D_{(l)} & 0\\ 0 & D^{(u)}\end{bmatrix}=\operatorname{diag}\left\{K1_n\right\}.$$

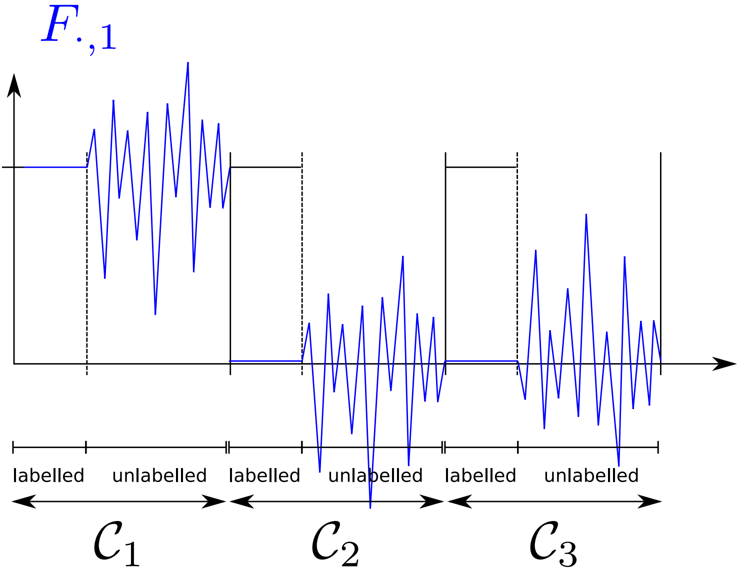


Figure: Typical expected performance output

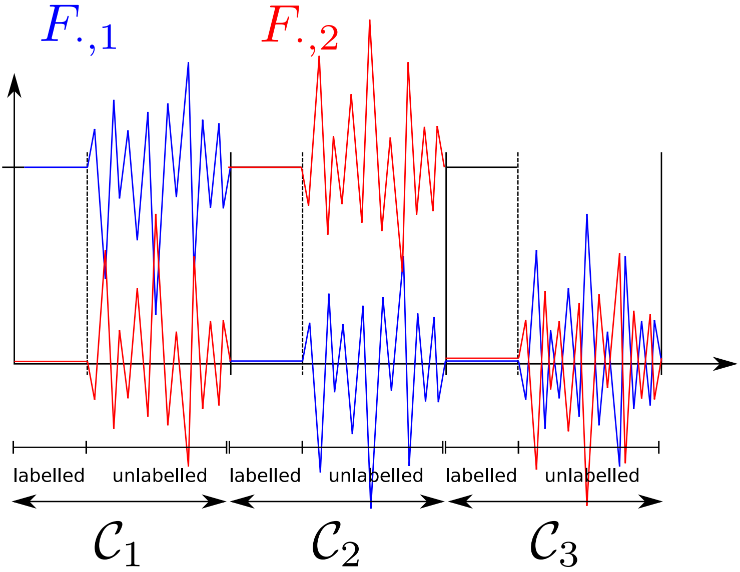


Figure: Typical expected performance output

The finite-dimensional intuition: What we expect

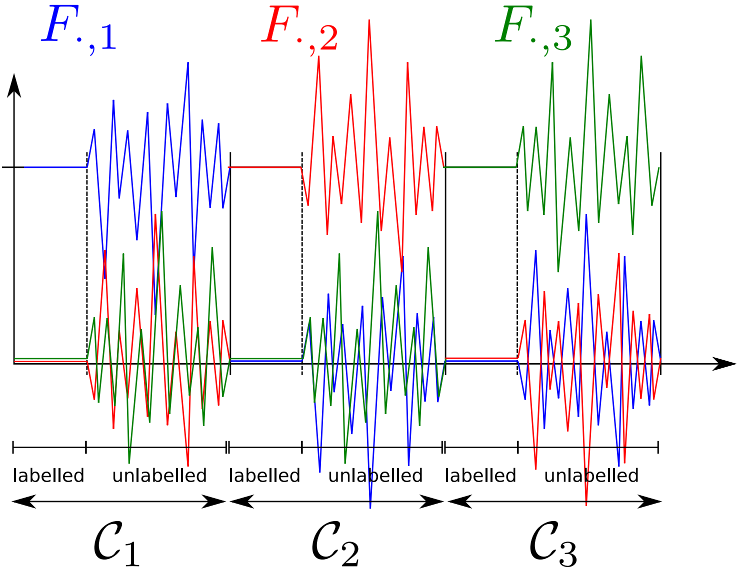


Figure: Typical expected performance output

MNIST Data Example

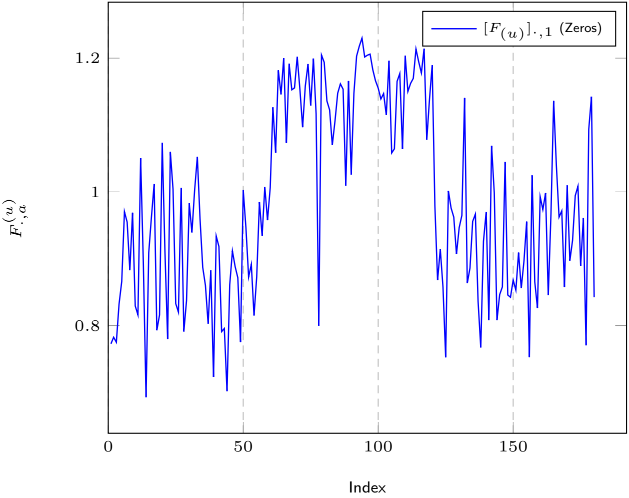


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

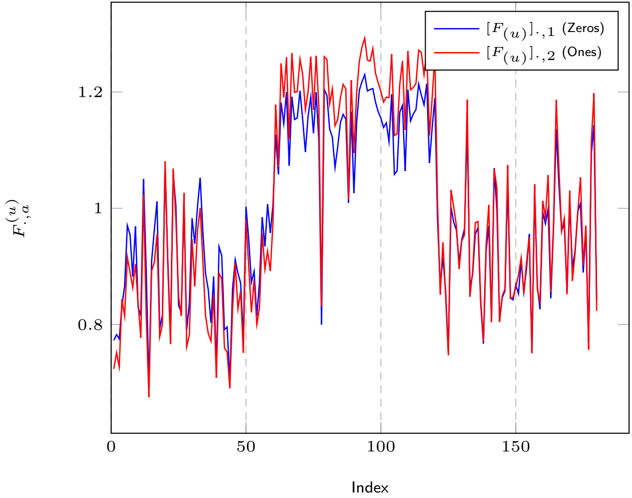


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

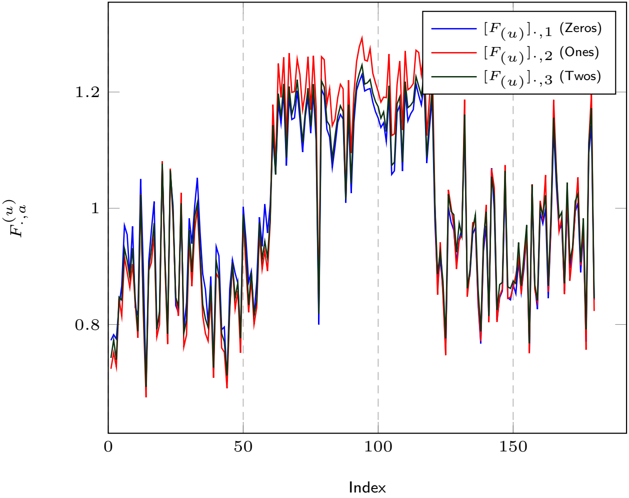


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

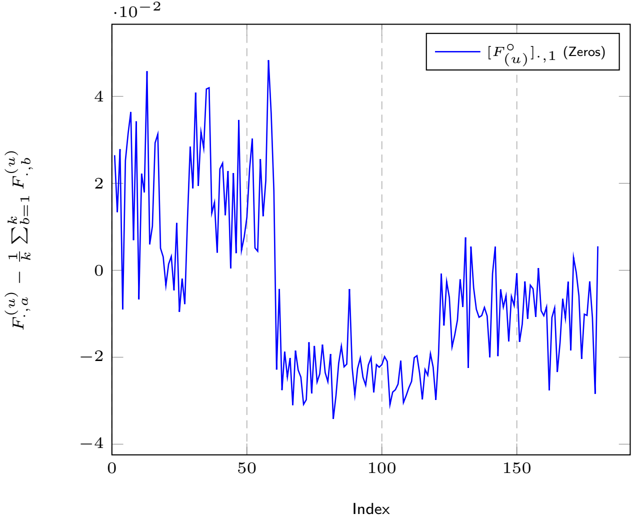


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k 1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

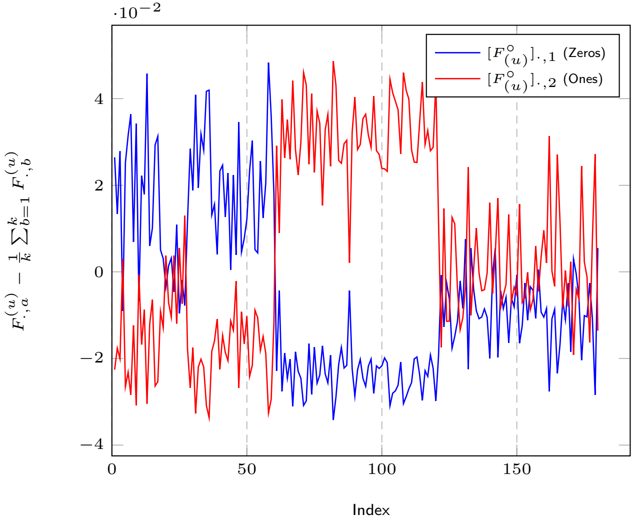


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

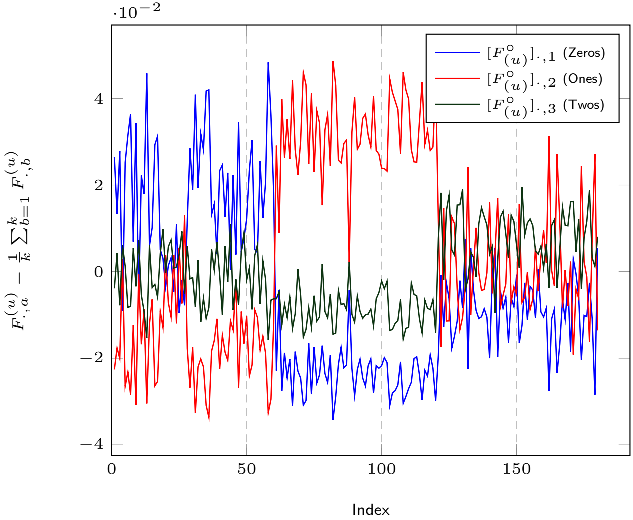


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k} F_{(u)} 1_k 1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Method: Assume $n_l/n \to c_l \in (0,1)$

- We aim at characterizing

$$F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1}\right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}$$

Method: Assume $n_l/n \to c_l \in (0,1)$

- 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 $n,p \to \infty$,

$$K_{(u,u)} = f(\tau) 1_{n_u} 1_{n_u}^{\mathsf{T}} + O_{\|\cdot\|}(n^{-\frac{1}{2}})$$
$$D_{(u)} = n f(\tau) I_{n_u} + O(n^{\frac{1}{2}})$$

and similarly for $K_{(u,l)}$, $D_{(l)}$.

Method: Assume $n_l/n \to c_l \in (0,1)$

- 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 $n, p \to \infty$,

$$K_{(u,u)} = f(\tau) 1_{n_u} 1_{n_u}^{\mathsf{T}} + O_{\|\cdot\|}(n^{-\frac{1}{2}})$$
$$D_{(u)} = n f(\tau) I_{n_u} + O(n^{\frac{1}{2}})$$

  and similarly for $K_{(u,l)}$, $D_{(l)}$.

- So that

$$\left(I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1}\right)^{-1} = \left(I_{n_u} - \frac{1_{n_u} 1_{n_u}^{\mathsf{T}}}{n} + O_{\|\cdot\|}(n^{-\frac{1}{2}})\right)^{-1}$$

  easily Taylor expanded.

Results: Assuming $n_l/n \to c_l \in (0,1)$, by previous Taylor expansion,

- In the first order,

$$F^{(u)}_{\cdot,a} = C\frac{n_{l,a}}{n}\left[\underbrace{v}_{O(1)} + \underbrace{\alpha\frac{t_a 1_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})}\right] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^{\circ}$.

Results: Assuming $n_l/n \to c_l \in (0,1)$, by previous Taylor expansion,

- In the first order,

$$F_{\cdot,a}^{(u)} = C\frac{n_{l,a}}{n}\left[\underbrace{v}_{O(1)} + \underbrace{\alpha\frac{t_a 1_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})}\right] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

  where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ$.

- Consequences:

Main Results

Results: Assuming $n_l/n \to c_l \in (0,1)$, by previous Taylor expansion,

- In the first order,

$$F_{\cdot,a}^{(u)} = C\frac{n_{l,a}}{n}\Big[\underbrace{v}_{O(1)} + \underbrace{\alpha\frac{t_a 1_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})}\Big] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

  where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ$.

- Consequences:
  - Random non-informative bias v

Main Results

Results: Assuming $n_l/n \to c_l \in (0,1)$, by previous Taylor expansion,

- In the first order,

$$F^{(u)}_{\cdot,a} = C\frac{n_{l,a}}{n}\Bigg[\underbrace{v}_{O(1)} + \alpha\underbrace{\frac{t_a 1_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})}\Bigg] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

  where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ$.

- Consequences:
  - Random non-informative bias v
  - Strong Impact of $n_{l,a}$

$$\boxed{F^{(u)}_{\cdot,a} \text{ to be scaled by } n_{l,a}}$$

Main Results

Results: Assuming $n_l/n \to c_l \in (0,1)$, by previous Taylor expansion,

- In the first order,

$$F^{(u)}_{\cdot,a} = C\frac{n_{l,a}}{n}\Big[\underbrace{v}_{O(1)} + \underbrace{\alpha\frac{t_a 1_{n_u}}{\sqrt{n}}}_{O(n^{-\frac{1}{2}})}\Big] + \underbrace{O(n^{-1})}_{\text{Informative terms}}$$

  where $v = O(1)$ random vector (entry-wise) and $t_a = \frac{1}{\sqrt{p}}\operatorname{tr} C^\circ_a$.

- Consequences:
  - Random non-informative bias v
  - Strong Impact of $n_{l,a}$

$$\boxed{F^{(u)}_{\cdot,a} \text{ to be scaled by } n_{l,a}}$$

  - Additional per-class bias $\alpha t_a 1_{n_u}$

$$\boxed{\alpha = 0 + \frac{\beta}{\sqrt{p}}.}$$

Main Results

As a consequence of the remarks above, we take

$$\alpha = \frac{\beta}{\sqrt{p}}$$

and define

$$\hat{F}_{i,a}^{(u)} = \frac{np}{n_{l,a}} F_{ia}^{(u)}.$$

As a consequence of the remarks above, we take

$$\alpha = \frac{\beta}{\sqrt{p}}$$

and define

$$\hat{F}^{(u)}_{i,a} = \frac{np}{n_{l,a}} F^{(u)}_{ia}.$$

Theorem

For $x_i \in \mathcal{C}_b$ unlabelled,

$$\boxed{\hat{F}_{i,\cdot} - G_b \to 0,\ G_b \sim \mathcal{N}(m_b, \Sigma_b)}$$

where $m_b \in \mathbb{R}^k$, $\Sigma_b \in \mathbb{R}^{k\times k}$ given by

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)}\tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)}\tilde{t}_a\tilde{t}_b + \frac{2f''(\tau)}{f(\tau)}\tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2}t_a t_b + \beta\frac{n}{n_l}\frac{f'(\tau)}{f(\tau)}t_a + B_b$$

$$(\Sigma_b)_{a_1a_2} = \frac{2trC_b^2}{p}\left(\frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)}\right)^2 t_{a_1}t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2}\left([M^{\mathsf{T}}C_bM]_{a_1a_2} + \frac{\delta_{a_1}^{a_2}p}{n_{l,a_1}}T_{ba_1}\right)$$

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.

Corollary (Asymptotic Classification Error)

For $k = 2$ classes and $a \neq b$,

$$P(\hat{F}_{i,a} > \hat{F}_{ib} \mid x_i \in \mathcal{C}_b) - Q\left(\frac{(m_b)_b - (m_b)_a}{\sqrt{[1,-1]\Sigma_b[1,-1]^{\mathsf{T}}}}\right) \to 0.$$

Corollary (Asymptotic Classification Error)

For $k = 2$ classes and $a \neq b$,

$$P(\hat{F}_{i,a} > \hat{F}_{ib} \mid x_i \in \mathcal{C}_b) - Q\left(\frac{(m_b)_b - (m_b)_a}{\sqrt{[1,-1]\Sigma_b[1,-1]^{\mathsf{T}}}}\right) \to 0.$$

Some consequences:

- non obvious choices of appropriate kernels
- non obvious choice of optimal β (induces a possibly beneficial bias)
- importance of n_l versus n_u.

MNIST Data Example

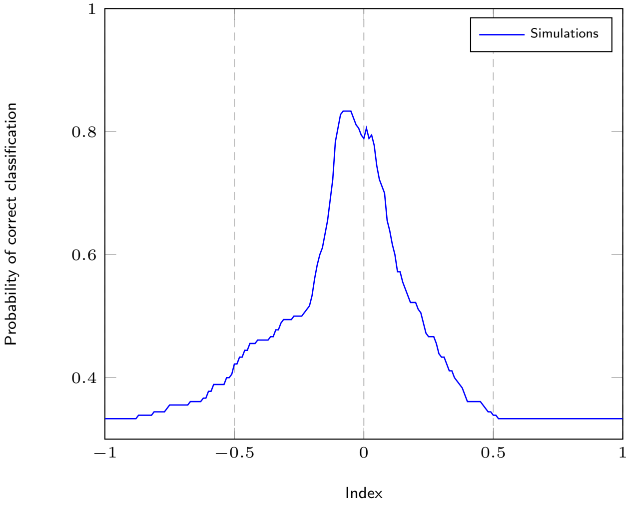


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

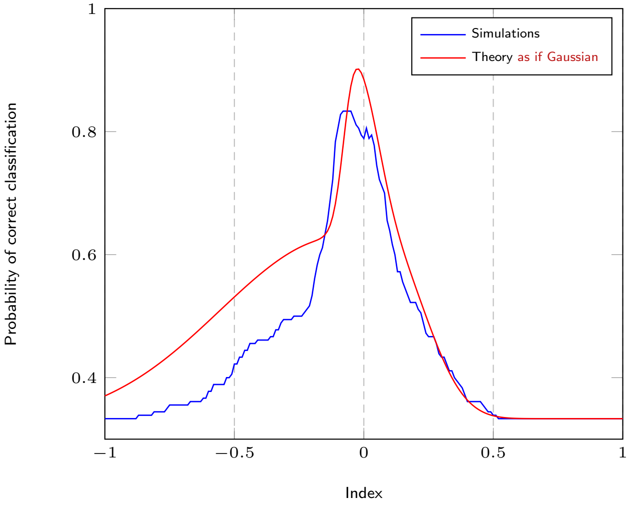


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

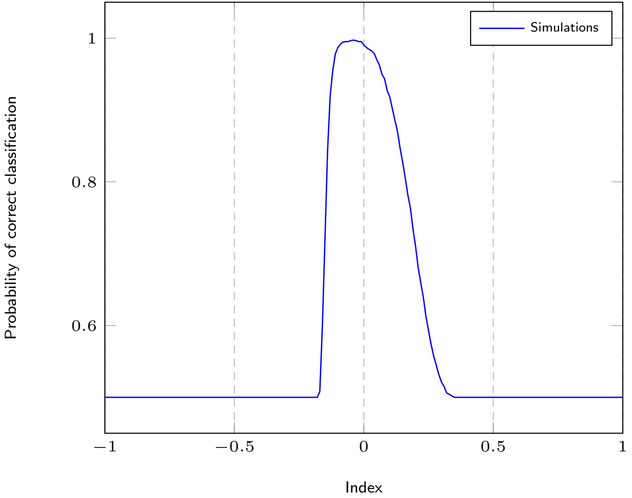


Figure: Performance as a function of α, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

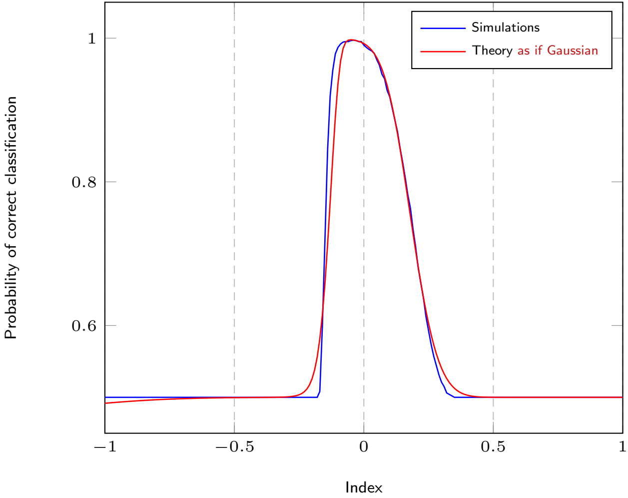


Figure: Performance as a function of α, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Reminder:
For $x_i \in \mathcal{C}_b$ unlabelled, $\hat{F}_{i,\cdot} - G_b \to 0$, $G_b \sim \mathcal{N}(m_b, \Sigma_b)$ with

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)}\tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)}\tilde{t}_a\tilde{t}_b + \frac{2f''(\tau)}{f(\tau)}\tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2}t_a t_b + \beta\frac{n}{n_l}\frac{f'(\tau)}{f(\tau)}t_a + B_b$$

$$(\Sigma_b)_{a_1 a_2} = \frac{2\mathrm{tr}\, C_b^2}{p}\left(\frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)}\right)^2 t_{a_1}t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2}\left([M^\mathsf{T} C_b M]_{a_1 a_2} + \frac{\delta_{a_1}^{a_2} p}{n_{l,a_1}}T_{ba_1}\right)$$

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.

Reminder:
For $x_i \in \mathcal{C}_b$ unlabelled, $\hat{F}_{i,\cdot} - G_b \to 0$, $G_b \sim \mathcal{N}(m_b, \Sigma_b)$ with

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)}\tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)}\tilde{t}_a\tilde{t}_b + \frac{2f''(\tau)}{f(\tau)}\tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2}t_a t_b + \beta\frac{n}{n_l}\frac{f'(\tau)}{f(\tau)}t_a + B_b$$

$$(\Sigma_b)_{a_1 a_2} = \frac{2\operatorname{tr} C_b^2}{p}\left(\frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)}\right)^2 t_{a_1}t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2}\left([M^{\mathsf{T}} C_b M]_{a_1 a_2} + \frac{\delta_{a_1}^{a_2} p}{n_{l,a_1}} T_{ba_1}\right)$$

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!
  $\longrightarrow$ increasing n_u asymptotically non beneficial.

Reminder:
For $x_i \in \mathcal{C}_b$ unlabelled, $\hat{F}_{i,\cdot} - G_b \to 0$, $G_b \sim \mathcal{N}(m_b, \Sigma_b)$ with

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)}\tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)}\tilde{t}_a\tilde{t}_b + \frac{2f''(\tau)}{f(\tau)}\tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2}t_a t_b + \beta\frac{n}{n_l}\frac{f'(\tau)}{f(\tau)}t_a + B_b$$

$$(\Sigma_b)_{a_1a_2} = \frac{2\mathrm{tr}\,C_b^2}{p}\left(\frac{f'(\tau)^2}{f(\tau)^2} - \frac{f''(\tau)}{f(\tau)}\right)^2 t_{a_1}t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2}\left([M^\mathsf{T}C_bM]_{a_1a_2} + \frac{\delta_{a_1}^{a_2}p}{n_{l,a_1}}T_{ba_1}\right)$$

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!
  $\longrightarrow$ increasing n_u asymptotically non beneficial.
- Even best Laplacian regularizer **brings SSL to be merely supervised learning**.

Outline

Resurrecting SSL by centering

Reminder:

$$F = \operatorname{argmin}_{F\in\mathbb{R}^{n\times k}} \sum_{a=1}^{k}\sum_{i,j} K_{ij}(F_{ia}d_i^{\alpha-1} - F_{ja}d_j^{\alpha-1})^2 \quad \text{with } F_{ia}^{(l)} = \boldsymbol{\delta}_{\{x_i\in\mathcal{C}_a\}}$$

$$\Leftrightarrow F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha}K_{(u,u)}D_{(u)}^{\alpha-1}\right)^{-1} D_{(u)}^{-\alpha}K_{(u,l)}D_{(l)}^{\alpha-1}F^{(l)}.$$

Reminder:

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^{k} \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2 \quad \text{with } F_{ia}^{(l)} = \boldsymbol{\delta}_{\{x_i \in \mathcal{C}_a\}}$$

$$\Leftrightarrow F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha} K_{(u,u)} D_{(u)}^{\alpha-1} \right)^{-1} D_{(u)}^{-\alpha} K_{(u,l)} D_{(l)}^{\alpha-1} F^{(l)}.$$

Domination of score flattening:

- Finite-dimensional intuition imposes K_{ij} decreasing with $\|x_i - x_j\| \Rightarrow$ solutions F_{ia} tend to "flatten"

Resurrecting SSL by centering

Reminder:

$$F = \operatorname{argmin}_{F\in\mathbb{R}^{n\times k}} \sum_{a=1}^{k}\sum_{i,j} K_{ij}(F_{ia}d_i^{\alpha-1} - F_{ja}d_j^{\alpha-1})^2 \quad \text{with } F_{ia}^{(l)} = \boldsymbol{\delta}_{\{x_i\in\mathcal{C}_a\}}$$

$$\Leftrightarrow F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha}K_{(u,u)}D_{(u)}^{\alpha-1}\right)^{-1} D_{(u)}^{-\alpha}K_{(u,l)}D_{(l)}^{\alpha-1}F^{(l)}.$$

Domination of score flattening:

- Finite-dimensional intuition imposes K_{ij} decreasing with $\|x_i - x_j\| \Rightarrow$ solutions F_{ia} tend to "flatten"
- **Consequence**: $D_{(u)}^{-\alpha}K_{(u,u)}D_{(u)}^{\alpha-1} \simeq \frac{1}{n}1_{n_u}1_{n_u}^{\mathsf{T}}$ and **clustering information vanishes** (not so obvious but can be shown).

Resurrecting SSL by centering

Reminder:

$$F = \operatorname{argmin}_{F\in\mathbb{R}^{n\times k}} \sum_{a=1}^{k}\sum_{i,j} K_{ij}(F_{ia}d_i^{\alpha-1} - F_{ja}d_j^{\alpha-1})^2 \quad \text{with } F_{ia}^{(l)} = \boldsymbol{\delta}_{\{x_i\in\mathcal{C}_a\}}$$

$$\Leftrightarrow F^{(u)} = \left(I_{n_u} - D_{(u)}^{-\alpha}K_{(u,u)}D_{(u)}^{\alpha-1}\right)^{-1} D_{(u)}^{-\alpha}K_{(u,l)}D_{(l)}^{\alpha-1}F^{(l)}.$$

Domination of score flattening:

- Finite-dimensional intuition imposes K_{ij} decreasing with $\|x_i - x_j\| \Rightarrow$ solutions F_{ia} tend to "flatten"
- **Consequence**: $D_{(u)}^{-\alpha}K_{(u,u)}D_{(u)}^{\alpha-1} \simeq \frac{1}{n}1_{n_u}1_{n_u}^{\mathsf{T}}$ and **clustering information vanishes** (not so obvious but can be shown).

Solution:

- Forgetting finite-dimensional intuition: **"recenter" K to kill flattening**, i.e., use

$$\boxed{\tilde{K} = PKP}, \; P = I_n - \frac{1}{n}1_n1_n^{\mathsf{T}}.$$

Theoretical results

Setting

- $K = 2$, $x_i \sim \mathcal{N}(\pm\mu, I_p)$
- scores $f_u = (\alpha I_{n_u} - \tilde{K}_{uu})^{-1} \tilde{K}_{ul} f_l$.

Setting

- $K = 2$, $x_i \sim \mathcal{N}(\pm\mu, I_p)$
- scores $f_u = (\alpha I_{n_u} - \tilde{K}_{uu})^{-1} \tilde{K}_{ul} f_l$.

Theorem (Asymptotic mean and variance)

As $n \to \infty$,

$$\frac{j_i^{(u)\mathsf{T}} f_u}{n_{ui}} - m_i \xrightarrow{\text{a.s.}} 0, \quad \frac{\left(f_u - m_i 1_{n_u}\right)^{\mathsf{T}} D_i^{(u)} \left(f_u - m_i 1_{n_u}\right)}{n_{ui}} - \sigma_i^2 \xrightarrow{\text{a.s.}} 0$$

where, for $i = 1, 2$,

$$m_i \equiv -\frac{c_l}{c_u} s_i \left(1 - \left[1 + \frac{c_u c_1 c_2 \|\mu\|^2}{c_0} \frac{\delta}{1+\delta}\right]^{-1}\right)$$

$$\sigma_i^2 \equiv \frac{s_i^2 c_l^2 c_i^2 \|\mu\|^2 \delta^2}{c_0^2 (1+\delta)^2 - c_u c_0 \delta^2} \frac{1 + \frac{c_u c_1 c_2 \|\mu\|^2}{c_0} \frac{\delta^2}{(1+\delta)^2}}{\left(1 + \frac{c_u c_1 c_2 \|\mu\|^2}{c_0} \frac{\delta}{1+\delta}\right)^2} + \frac{s_i^2 c_l c_i}{1 - c_i} \frac{\delta^2}{c_0 (1+\delta)^2 - c_u \delta^2}$$

with δ *defined as*

$$\delta \equiv -\frac{1}{2} + \frac{c_u - c_0 + \operatorname{sign}(\alpha)\sqrt{(\alpha - \alpha_-)(\alpha - \alpha_+)}}{2\alpha}.$$

Performance as a function of n_u, n_l

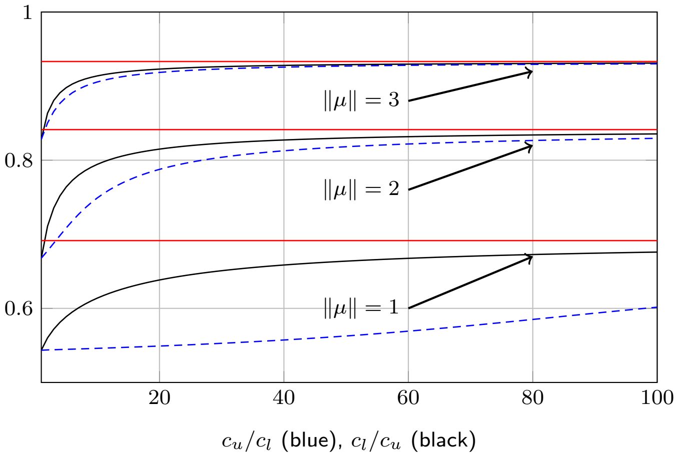


Figure: Correct classification rate, at optimal α, as a function of (i) n_u for fixed $p/n_l = 5$ (blue) and (ii) n_l for fixed $p/n_u = 5$ (black); $c_1 = c_2 = \frac{1}{2}$; different values for $\|\mu\|$. Comparison to optimal Neyman–Pearson performance for known μ (in red).

The spike case or not (1)

Marčenko–Pastur + spike limit

- limiting eigenvalue distribution is Marčenko–Pastur law

The spike case or not (1)

Marčenko–Pastur + spike limit

- limiting eigenvalue distribution is Marčenko–Pastur law
- presence of **isolated spike iif**

$$\boxed{\|\mu\|^2 > \frac{1}{c_1 c_2}\sqrt{\frac{c_0}{c_u}}.}$$

The spike case or not (1)

Marčenko–Pastur + spike limit

- limiting eigenvalue distribution is Marčenko–Pastur law
- presence of **isolated spike iif**

$$\boxed{\|\mu\|^2 > \frac{1}{c_1 c_2}\sqrt{\frac{c_0}{c_u}}.}$$

- determines **existence or not of unsupervised spectral clustering solution**.

The spike case or not (1)

Marčenko–Pastur + spike limit

- limiting eigenvalue distribution is Marčenko–Pastur law
- presence of **isolated spike iif**

$$\boxed{\|\mu\|^2 > \frac{1}{c_1 c_2}\sqrt{\frac{c_0}{c_u}}.}$$

- determines **existence or not of unsupervised spectral clustering solution**.

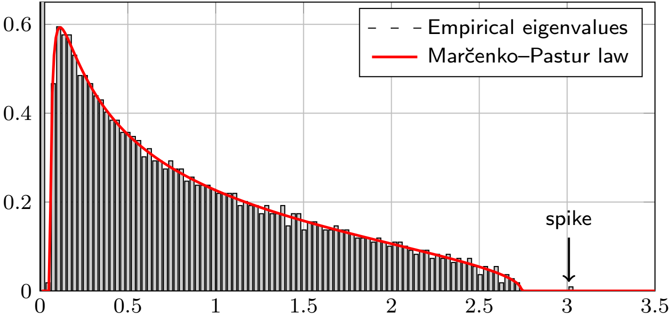


Figure: Eigenvalue distribution of K_{uu} versus the (scaled) Marčenko–Pastur law with Stieltjes transform δ, for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$. The value $\|\mu\| = 2.5$ ensures the presence of a leading isolated eigenvalue (spike).

The spike case or not (2)

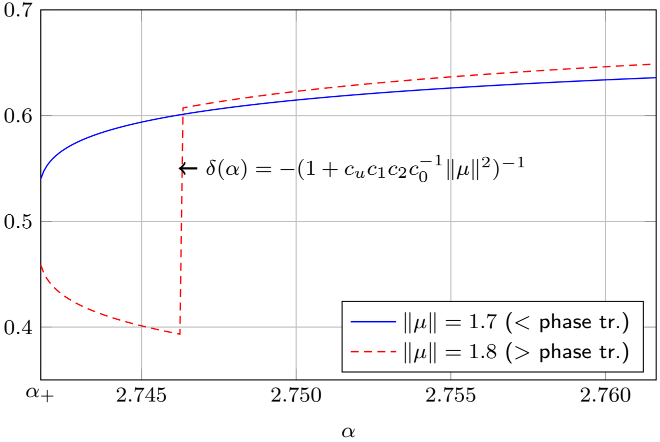


Figure: Asymptotic correct classification probability $\Phi\left(\frac{m_1}{\sigma_1}\right)$ as a function of α for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$, $c_1 = \frac{1}{2}$, two different values of $\|\mu\|$, below and above phase transition.

SSL: the road from supervised to unsupervised

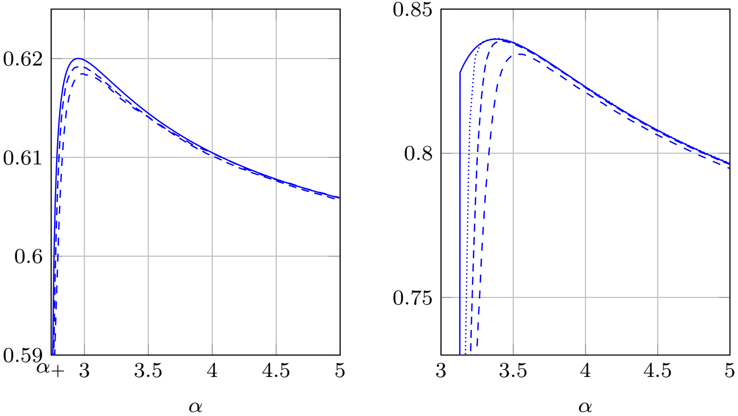


Figure: Theory (solid) versus practice (dashed; from right to left: $n = 400, 1000, 4000$): correct classification probability as a function of α for $c_u = \frac{9}{10}$, $c_0 = \frac{1}{2}$, $c_1 = \frac{1}{2}$, and **left**: $\|\mu\| = 1.5$ (below phase transition); **right**: $\|\mu\| = 2.5$ (above phase transition). Different values of n.

Experimental evidence: MNIST

Digits	(0,8)	(2,7)	(6,9)
		$n_u = 100$	
Centered kernel	**89.5±3.6**	**89.5±3.4**	**85.3±5.9**
Iterated centered kernel	**89.5±3.6**	**89.5±3.4**	**85.3±5.9**
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	92.2±0.9	92.5±0.8	92.6±1.6
Iterated centered kernel	**92.3±0.9**	**92.5± 0.8**	**92.9±1.4**
Laplacian	65.6±4.1	74.4±4.0	69.5±3.7
Iterated Laplacian	**92.2±0.9**	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)

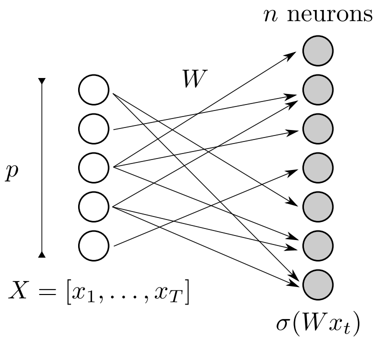
Class ID	(2,7)	(9,10)	(11,18)
	$n_u = 100$		
Centered kernel	79.0±10.4	77.5±9.2	78.5±7.1
Iterated centered kernel	**85.3±5.9**	**89.2±5.6**	**90.1±6.7**
Laplacian	73.8±9.8	77.3±9.5	78.6±7.2
Iterated Laplacian	83.7±7.2	88.0±6.8	87.1±8.8
Manifold	77.6±8.9	81.4±10.4	82.3±10.8
	$n_u = 1000$		
Centered kernel	83.6±2.4	84.6±2.4	88.7±9.4
Iterated centered kernel	**84.8±3.8**	**88.0±5.5**	**96.4±3.0**
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 $n_u = 1000$.

Outline

Context: Random Feature Map

- (large) input $x_1, \ldots, x_T \in \mathbb{R}^p$
- random $W = \begin{bmatrix} w_1^\mathsf{T} \\ \ldots \\ w_n^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{n \times p}$
- non-linear activation function σ.

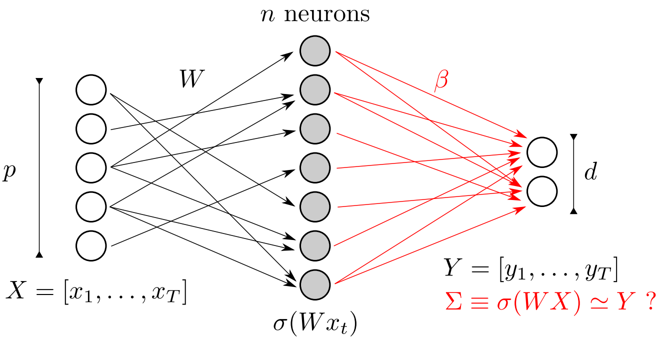


Context: Random Feature Map

- (large) input $x_1, \ldots, x_T \in \mathbb{R}^p$
- random $W = \begin{bmatrix} w_1^\mathsf{T} \\ \ldots \\ w_n^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{n \times p}$
- non-linear activation function σ.

Neural Network Model (extreme learning machine): Ridge-regression learning

- small output $y_1, \ldots, y_T \in \mathbb{R}^d$
- ridge-regression output $\beta \in \mathbb{R}^{n \times d}$



Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

- **Training MSE**:

$$E_{\mathrm{train}} = \frac{1}{T} \sum_{i=1}^{T} \|y_i - \beta^{\mathsf{T}} \sigma(W x_i)\|^2 = \frac{1}{T} \|Y - \beta^{\mathsf{T}} \Sigma\|_F^2$$

with

$$\Sigma = \sigma(WX) = \left\{ \sigma(w_i^{\mathsf{T}} x_j) \right\}_{\substack{1 \leq i \leq n \\ 1 \leq j \leq T}}$$

$$\beta = \frac{1}{T} \Sigma \left(\frac{1}{T} \Sigma^{\mathsf{T}} \Sigma + \gamma I_T \right)^{-1} Y.$$

Objectives: evaluate training and testing MSE performance as $n, p, T \to \infty$

- **Training MSE**:

$$E_{\text{train}} = \frac{1}{T} \sum_{i=1}^{T} \|y_i - \beta^{\mathsf{T}} \sigma(W x_i)\|^2 = \frac{1}{T} \|Y - \beta^{\mathsf{T}} \Sigma\|_F^2$$

with

$$\Sigma = \sigma(WX) = \left\{ \sigma(w_i^{\mathsf{T}} x_j) \right\}_{\substack{1 \le i \le n \\ 1 \le j \le T}}$$

$$\beta = \frac{1}{T} \Sigma \left(\frac{1}{T} \Sigma^{\mathsf{T}} \Sigma + \gamma I_T \right)^{-1} Y.$$

- **Testing MSE**: upon new pair $(\hat{X}, \hat{Y})$ of length $\hat{T}$,

$$E_{\text{test}} = \frac{1}{\hat{T}} \|\hat{Y} - \beta^{\mathsf{T}} \hat{\Sigma}\|_F^2.$$

where $\hat{\Sigma} = \sigma(W\hat{X})$.

Preliminary observations:

- Link to resolvent of $\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma$:

$$E_{\text{train}} = \frac{\gamma^2}{T}\operatorname{tr} Y^{\mathsf{T}} Y Q^2 = -\gamma^2 \frac{\partial}{\partial\gamma}\frac{1}{T}\operatorname{tr} Y^{\mathsf{T}} Y Q$$

where $Q = Q(\gamma)$ is the resolvent

$$Q \equiv \left(\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma + \gamma I_T\right)^{-1}$$

with $\Sigma_{ij} = \sigma(w_i^{\mathsf{T}} x_j)$.

Preliminary observations:

- Link to resolvent of $\frac{1}{T}\Sigma^\mathsf{T}\Sigma$:

$$E_{\text{train}} = \frac{\gamma^2}{T} \operatorname{tr} Y^\mathsf{T} Y Q^2 = -\gamma^2 \frac{\partial}{\partial \gamma} \frac{1}{T} \operatorname{tr} Y^\mathsf{T} Y Q$$

where $Q = Q(\gamma)$ is the resolvent

$$Q \equiv \left(\frac{1}{T}\Sigma^\mathsf{T}\Sigma + \gamma I_T \right)^{-1}$$

with $\Sigma_{ij} = \sigma(w_i^\mathsf{T} x_j)$.

$\boxed{\text{Central object: resolvent } E[Q].}$

Main Technical Result

Theorem [Asymptotic Equivalent for $E[Q]$]

For Lipschitz σ, bounded $\|X\|, \|Y\|$, $W = f(Z)$ (entry-wise) with Z standard Gaussian, we have, for all $\varepsilon > 0$,

$$\left\| E[Q] - \bar{Q} \right\| < C n^{\varepsilon - \frac{1}{2}}$$

for some $C > 0$, where

$$\bar{Q} = \left(\frac{n}{T} \frac{\Phi}{1 + \delta} + \gamma I_T \right)^{-1}$$

$$\Phi \equiv E\left[\sigma(X^\mathsf{T} w) \sigma(w^\mathsf{T} X) \right]$$

with $w = f(z)$, $z \sim \mathcal{N}(0, I_p)$, and $\delta > 0$ the unique positive solution to

$$\delta = \frac{1}{T} \operatorname{tr} \Phi \bar{Q}.$$

Main Technical Result

Theorem [Asymptotic Equivalent for $E[Q]$]

For Lipschitz σ, bounded $\|X\|, \|Y\|$, $W = f(Z)$ (entry-wise) with Z standard Gaussian, we have, for all $\varepsilon > 0$,

$$\left\| E[Q] - \bar{Q} \right\| < C n^{\varepsilon - \frac{1}{2}}$$

for some $C > 0$, where

$$\bar{Q} = \left(\frac{n}{T} \frac{\Phi}{1+\delta} + \gamma I_T \right)^{-1}$$

$$\Phi \equiv E\left[\sigma(X^\mathsf{T} w) \sigma(w^\mathsf{T} X) \right]$$

with $w = f(z)$, $z \sim \mathcal{N}(0, I_p)$, and $\delta > 0$ the unique positive solution to

$$\delta = \frac{1}{T} \operatorname{tr} \Phi \bar{Q}.$$

Proof arguments:

- $\sigma(WX)$ has independent rows but dependent columns
- breaks the "trace lemma" argument (i.e., $\frac{1}{p} w^\mathsf{T} X A X^\mathsf{T} w \simeq \frac{1}{p} \operatorname{tr} X A X^\mathsf{T}$)

Theorem [Asymptotic Equivalent for $E[Q]$]

For Lipschitz σ, bounded $\|X\|, \|Y\|$, $W = f(Z)$ (entry-wise) with Z standard Gaussian, we have, for all $\varepsilon > 0$,

$$\left\|E[Q] - \bar{Q}\right\| < Cn^{\varepsilon - \frac{1}{2}}$$

for some $C > 0$, where

$$\bar{Q} = \left(\frac{n}{T} \frac{\Phi}{1+\delta} + \gamma I_T \right)^{-1}$$

$$\Phi \equiv E\left[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)\right]$$

with $w = f(z)$, $z \sim \mathcal{N}(0, I_p)$, and $\delta > 0$ the unique positive solution to

$$\delta = \frac{1}{T}\operatorname{tr} \Phi\bar{Q}.$$

Proof arguments:

- $\sigma(WX)$ has independent rows but dependent columns
- breaks the "trace lemma" argument (i.e., $\frac{1}{p} w^\mathsf{T} X A X^\mathsf{T} w \simeq \frac{1}{p} \operatorname{tr} XAX^\mathsf{T}$)

$$\boxed{\text{Concentration of measure: } P\left(\left| \frac{1}{p}\sigma(w^\mathsf{T} X)A\sigma(X^\mathsf{T} w) - \frac{1}{p}\operatorname{tr} \Phi A \right| > t \right) \leq Ce^{-cn\min(t,t^2)}}$$

- Values of $\Phi(a,b)$ for $w \sim \mathcal{N}(0, I_p)$,

$\sigma(t)$	$\Phi(a,b)$
$\max(t,0)$	$\frac{1}{2\pi}\|a\|\|b\|\left(\angle(a,b)\,\mathrm{acos}(-\angle(a,b)) + \sqrt{1-\angle(a,b)^2}\right)$
$\lvert t\rvert$	$\frac{2}{\pi}\|a\|\|b\|\left(\angle(a,b)\,\mathrm{asin}(\angle(a,b)) + \sqrt{1-\angle(a,b)^2}\right)$
$\mathrm{erf}(t)$	$\frac{2}{\pi}\,\mathrm{asin}\left(\frac{2a^\mathsf{T}b}{\sqrt{(1+2\|a\|^2)(1+2\|b\|^2)}}\right)$
$1_{\{t>0\}}$	$\frac{1}{2} - \frac{1}{2\pi}\,\mathrm{acos}(\angle(a,b))$
$\mathrm{sign}(t)$	$1 - \frac{2}{\pi}\,\mathrm{acos}(\angle(a,b))$
$\cos(t)$	$\exp(-\frac{1}{2}(\|a\|^2 + \|b\|^2))\cosh(a^\mathsf{T}b)$.

where $\angle(a,b) \equiv \frac{a^\mathsf{T}b}{\|a\|\|b\|}$.

- Values of $\Phi(a,b)$ for $w \sim \mathcal{N}(0, I_p)$,

$\sigma(t)$	$\Phi(a,b)$
$\max(t,0)$	$\frac{1}{2\pi}\|a\|\|b\|\left(\angle(a,b)\,\mathrm{acos}(-\angle(a,b)) + \sqrt{1-\angle(a,b)^2}\right)$
$\lvert t\rvert$	$\frac{2}{\pi}\|a\|\|b\|\left(\angle(a,b)\,\mathrm{asin}(\angle(a,b)) + \sqrt{1-\angle(a,b)^2}\right)$
$\mathrm{erf}(t)$	$\frac{2}{\pi}\,\mathrm{asin}\left(\frac{2a^\mathsf{T} b}{\sqrt{(1+2\|a\|^2)(1+2\|b\|^2)}}\right)$
$1_{\{t>0\}}$	$\frac{1}{2} - \frac{1}{2\pi}\,\mathrm{acos}(\angle(a,b))$
$\mathrm{sign}(t)$	$1 - \frac{2}{\pi}\,\mathrm{acos}(\angle(a,b))$
$\cos(t)$	$\exp(-\frac{1}{2}(\|a\|^2+\|b\|^2))\cosh(a^\mathsf{T} b)$.

where $\angle(a,b) \equiv \frac{a^\mathsf{T} b}{\|a\|\|b\|}$.

- Value of $\Phi(a,b)$ for w_i i.i.d. with $E[w_i^k] = m_k$ $(m_1 = 0)$, $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$

$$\Phi(a,b) = \zeta_2^2\left[m_2^2\left(2(a^\mathsf{T} b)^2 + \|a\|^2\|b\|^2\right) + (m_4 - 3m_2^2)(a^2)^\mathsf{T}(b^2)\right] + \zeta_1^2 m_2 a^\mathsf{T} b$$
$$+ \zeta_2\zeta_1 m_3\left[(a^2)^\mathsf{T} b + a^\mathsf{T}(b^2)\right] + \zeta_2\zeta_0 m_2\left[\|a\|^2 + \|b\|^2\right] + \zeta_0^2$$

where $(a^2) \equiv [a_1^2, \ldots, a_p^2]^\mathsf{T}$.

Theorem [Asymptotic E_{train}]

For all $\varepsilon > 0$,

$$n^{\frac{1}{2}-\varepsilon}\left(E_{\text{train}} - \bar{E}_{\text{train}}\right) \to 0$$

almost surely, where

$$E_{\text{train}} = \frac{1}{T}\left\|Y^{\mathsf{T}} - \Sigma^{\mathsf{T}}\beta\right\|_F^2 = \frac{\gamma^2}{T}\operatorname{tr} Y^{\mathsf{T}} Y Q^2$$

$$\bar{E}_{\text{train}} = \frac{\gamma^2}{T}\operatorname{tr} Y^{\mathsf{T}} Y \bar{Q}\left[\frac{\frac{1}{n}\operatorname{tr}\Psi\bar{Q}^2}{1-\frac{1}{n}\operatorname{tr}(\Psi\bar{Q})^2}\Psi + I_T\right]\bar{Q}$$

with $\Psi \equiv \frac{n}{T}\frac{\Phi}{1+\delta}$.

- Letting $\hat{X} \in \mathbb{R}^{p \times \hat{T}}$, $\hat{Y} \in \mathbb{R}^{d \times \hat{T}}$ satisfy "similar properties" as (X, Y),

Claim [Asymptotic $E_{\rm test}$]

For all $\varepsilon > 0$,

$$n^{\frac{1}{2}-\varepsilon}\left(E_{\rm test} - \bar{E}_{\rm test}\right) \to 0$$

almost surely, where

$$\begin{aligned}
E_{\rm test} &= \frac{1}{\hat{T}} \left\| \hat{Y}^{\mathsf{T}} - \hat{\Sigma}^{\mathsf{T}} \beta \right\|_F^2 \\
\bar{E}_{\rm test} &= \frac{1}{\hat{T}} \left\| \hat{Y}^{\mathsf{T}} - \Psi_{X\hat{X}}^{\mathsf{T}} \bar{Q} Y^{\mathsf{T}} \right\|_F^2 \\
&\quad + \frac{\frac{1}{n} \operatorname{tr} Y^{\mathsf{T}} Y \bar{Q} \Psi \bar{Q}}{1 - \frac{1}{n} \operatorname{tr} (\Psi \bar{Q})^2} \left[\frac{1}{\hat{T}} \operatorname{tr} \Psi_{\hat{X}\hat{X}} - \frac{1}{\hat{T}} \operatorname{tr} (I_T + \gamma \bar{Q})(\Psi_{X\hat{X}} \Psi_{\hat{X}X} \bar{Q}) \right]
\end{aligned}$$

with $\Psi_{AB} = \frac{n}{T} \frac{\Phi_{AB}}{1+\delta}$, $\Phi_{AB} = E[\sigma(A^{\mathsf{T}} w) \sigma(w^{\mathsf{T}} B)]$.

Simulations on MNIST: Lipschitz $\sigma(\cdot)$

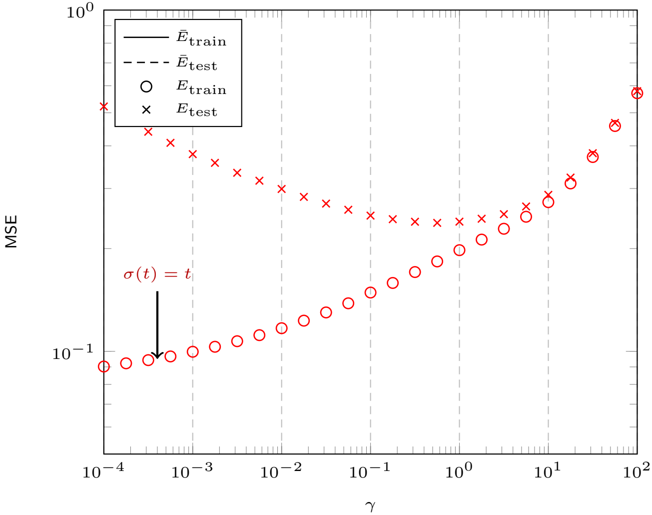


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations on MNIST: Lipschitz $\sigma(\cdot)$

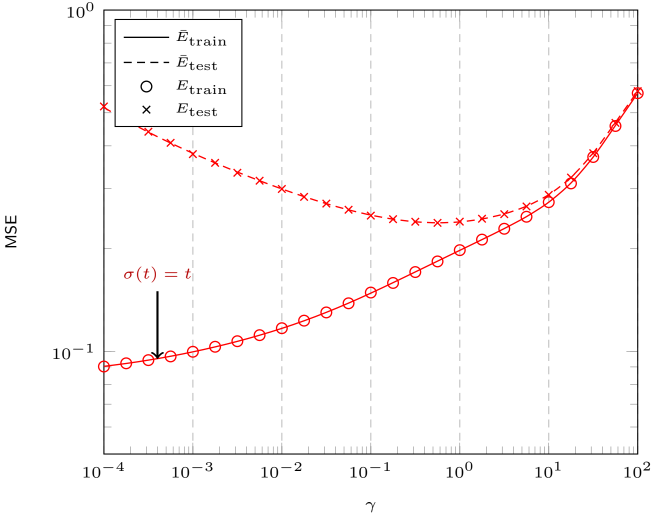


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

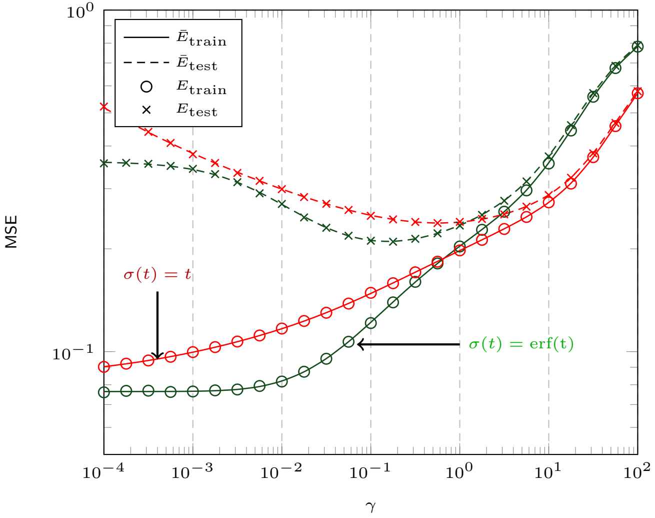


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations on MNIST: Lipschitz $\sigma(\cdot)$

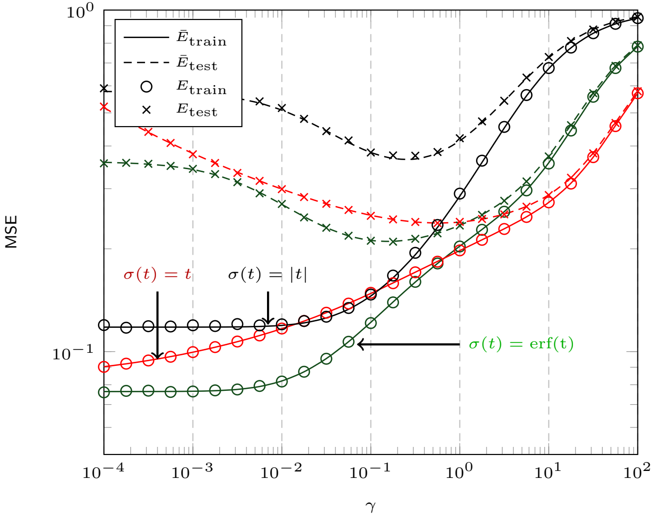


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations on MNIST: Lipschitz $\sigma(\cdot)$

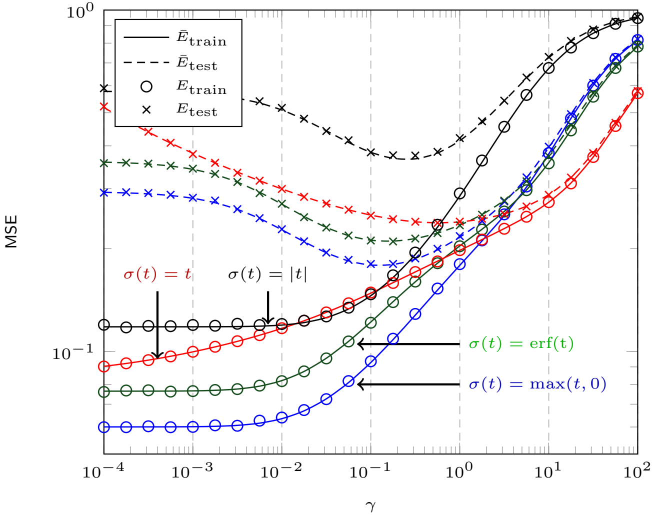


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations on MNIST: non Lipschitz $\sigma(\cdot)$

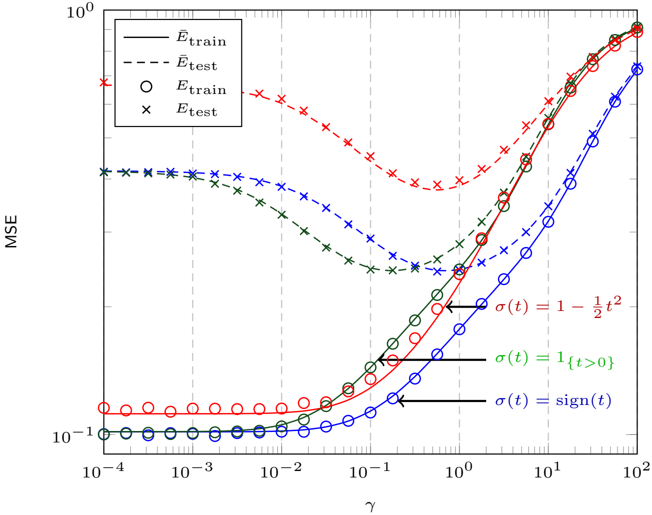


Figure: Neural network performance for $\sigma(\cdot)$ either discontinuous or non Lipschitz, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Deeper investigation on Φ

Statistical Assumptions on X

- Gaussian mixture model

$$x_i \in \mathcal{C}_a \Leftrightarrow x_i \sim \mathcal{N}(\frac{1}{\sqrt{p}}\mu_a, \frac{1}{p}C_a).$$

- **Growth rate**: $\|\mu_a^\circ\| = O(1)$, $\frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ = O(1)$.

Statistical Assumptions on X

- Gaussian mixture model

$$x_i \in \mathcal{C}_a \Leftrightarrow x_i \sim \mathcal{N}(\frac{1}{\sqrt{p}}\mu_a, \frac{1}{p}C_a).$$

- **Growth rate**: $\|\mu_a^\circ\| = O(1)$, $\frac{1}{\sqrt{p}}\operatorname{tr} C_a^\circ = O(1)$.

Theorem

As $p, T \to \infty$, for all $\sigma(\cdot)$ given in next table,

$$\|P\Phi P - P\tilde{\Phi}P\| \xrightarrow{\text{a.s.}} 0$$

with

$$\begin{aligned}
\tilde{\Phi} &\equiv d_1 \left(\Omega + M\frac{J^\mathsf{T}}{\sqrt{p}}\right)^\mathsf{T} \left(\Omega + M\frac{J^\mathsf{T}}{\sqrt{p}}\right) + d_2 UBU^\mathsf{T} + d_0 I_T \\
U &\equiv \left[\frac{J}{\sqrt{p}}, \phi\right] \\
B &\equiv \begin{bmatrix} tt^\mathsf{T} + 2T & t \\ t^\mathsf{T} & 1 \end{bmatrix}
\end{aligned}$$

and d_0, d_1, d_2 given in next table ($\phi_i = \|w_i\|^2 - E[\|w_i\|^2]$ for $x_i = \frac{1}{\sqrt{p}}\mu_a + w_i$).

Figure: Coefficients d_i in $\tilde{\Phi}$ for different $\sigma(\cdot)$.

$\sigma(t)$	d_0	d_1	d_2
t	0	1	0
$\mathrm{ReLU}(t)$	$\left(\frac{1}{4}-\frac{1}{2\pi}\right)\tau$	$\frac{1}{4}$	$\frac{1}{8\pi\tau}$
$\|t\|$	$\left(1-\frac{2}{\pi}\right)\tau$	0	$\frac{1}{2\pi\tau}$
$\mathrm{LReLU}(t)$	$\frac{\pi-2}{4\pi}(\varsigma_+ + \varsigma_-)^2\tau$	$\frac{1}{4}(\varsigma_+ - \varsigma_-)^2$	$\frac{1}{8\tau\pi}(\varsigma_+ + \varsigma_-)^2$
$1_{t>0}$	$\frac{1}{4}-\frac{1}{2\pi}$	$\frac{1}{2\pi\tau}$	0
$\mathrm{sign}(t)$	$1-\frac{2}{\pi}$	$\frac{2}{\pi\tau}$	0
$\varsigma_2 t^2+\varsigma_1 t+\varsigma_0$	$2\tau^2\varsigma_2^2$	ς_1^2	ς_2^2
$\cos(t)$	$\frac{1}{2}+\frac{e^{-2\tau}}{2}-e^{-\tau}$	0	$\frac{e^{-\tau}}{4}$
$\sin(t)$	$\frac{1}{2}-\frac{e^{-2\tau}}{2}-\tau e^{-\tau}$	$e^{-\tau}$	0
$\mathrm{erf}(t)$	$\frac{2}{\pi}\left(\arccos\left(\frac{2\tau}{2\tau+1}\right)-\frac{2\tau}{2\tau+1}\right)$	$\frac{4}{\pi}\frac{1}{2\tau+1}$	0
$\exp(-\frac{t^2}{2})$	$\frac{1}{\sqrt{2\tau+1}}-\frac{1}{\tau+1}$	0	$\frac{1}{4(\tau+1)^3}$

where

- $\mathrm{ReLU}(t)=\max(t,0)$
- $\mathrm{LReLU}(t)=\varsigma_+\max(t,0)+\varsigma_-\max(-t,0)$.

Deeper investigation on Φ

Three groups of functions $\sigma(\cdot)$ emerge:

- "means-oriented": $d_2 = 0$
- "covariance-oriented": $d_1 = 0$
- "balanced": $d_1, d_2 \neq 0$

Three groups of functions $\sigma(\cdot)$ emerge:

- "means-oriented": $d_2 = 0$
- "covariance-oriented": $d_1 = 0$
- "balanced": $d_1, d_2 \neq 0$

Case of the Leaky–ReLU

- $\sigma(t) = \varsigma_+ \max(t, 0) + \varsigma_- \max(-t, 0)$

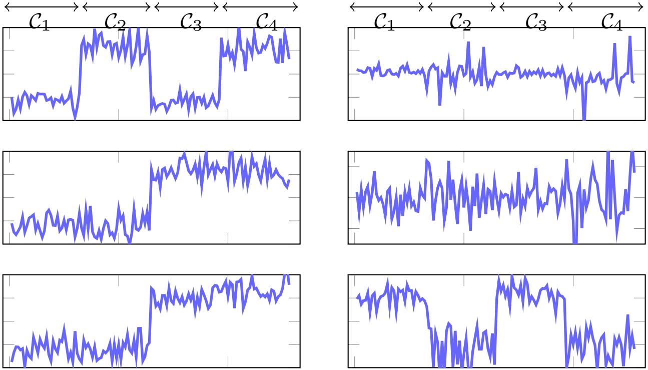
Deeper investigation on Φ

Three groups of functions $\sigma(\cdot)$ emerge:

- "means-oriented": $d_2 = 0$
- "covariance-oriented": $d_1 = 0$
- "balanced": $d_1, d_2 \neq 0$

Case of the Leaky–ReLU

- $\sigma(t) = \varsigma_+ \max(t, 0) + \varsigma_- \max(-t, 0)$



$\varsigma_+ = \varsigma_- = 1$

$\varsigma_+ = \varsigma_- = 1$

$\varsigma_+ = 1, \varsigma_- = 0$

Figure: Eigenvectors 1 and 2 of $P\Phi P$ for: $\mathcal{N}(\mu_1, C_1)$, $\mathcal{N}(\mu_1, C_2)$, $\mathcal{N}(\mu_2, C_1)$, $\mathcal{N}(\mu_2, C_2)$

Table: Clustering accuracies for different $\sigma(t)$ on MNIST dataset ($n = 32$).

	$\sigma(t)$	$T = 32$	$T = 64$	$T = 128$
MEAN-ORIENTED	t	85.31%	**88.94%**	87.30%
	$1_{t>0}$	86.00%	82.94%	85.56%
	$\mathrm{sign}(t)$	81.94%	83.34%	85.22%
	$\sin(t)$	85.31%	87.81%	**87.50%**
	$\mathrm{erf}(t)$	**86.50%**	87.28%	86.59%
COV-ORIENTED	$\|t\|$	62.81%	60.41%	57.81%
	$\cos(t)$	62.50%	59.56%	57.72%
	$\exp(-\frac{t^2}{2})$	64.00%	60.44%	58.67%
BALANCED	(t)	82.87%	85.72%	82.27%

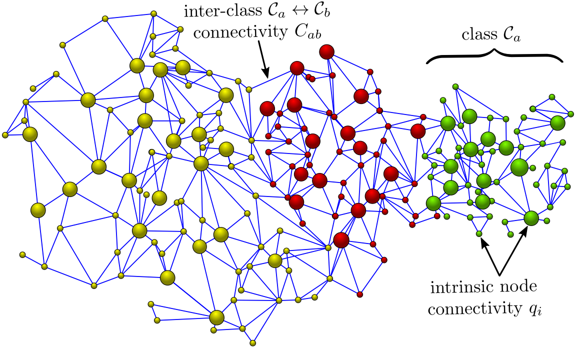
Depper investigation on Φ: Simulation results

Table: Clustering accuracies for different $\sigma(t)$ on epileptic EEG dataset ($n = 32$).

	$\sigma(t)$	$T = 32$	$T = 64$	$T = 128$
MEAN-ORIENTED	t	71.81%	70.31%	69.58%
	$1_{t>0}$	65.19%	65.87%	63.47%
	$\mathrm{sign}(t)$	67.13%	64.63%	63.03%
	$\sin(t)$	71.94%	70.34%	68.22%
	$\mathrm{erf}(t)$	69.44%	70.59%	67.70%
COV-ORIENTED	$\|t\|$	99.69%	99.69%	99.50%
	$\cos(t)$	99.00%	99.38%	99.36%
	$\exp(-\frac{t^2}{2})$	**99.81%**	**99.81%**	**99.77%**
BALANCED	(t)	84.50%	87.91%	90.97%

Outline

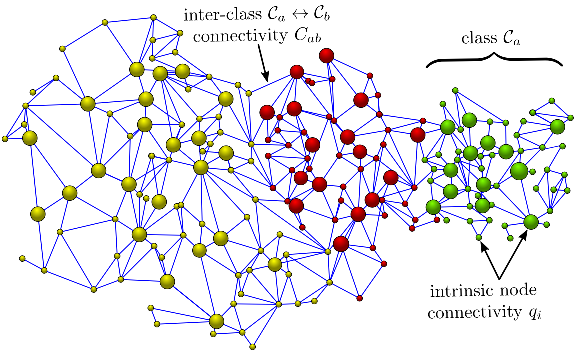
System Setting



Undirected graph with n nodes, m edges:

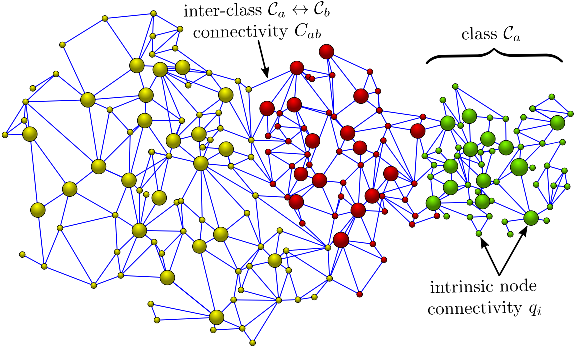
- "intrinsic" average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.

System Setting



Undirected graph with n nodes, m edges:

- "intrinsic" average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment C_{ab} between $\mathcal{C}_a$ and $\mathcal{C}_b$

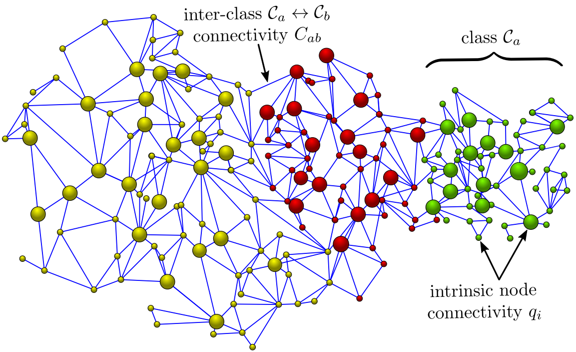


Undirected graph with n nodes, m edges:

- "intrinsic" average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment C_{ab} between $\mathcal{C}_a$ and $\mathcal{C}_b$
- edge probability for nodes $i \in \mathcal{C}_{g_i}$:

$$P(i \sim j) = q_i q_j C_{g_i g_j}.$$

System Setting



Undirected graph with n nodes, m edges:

- "intrinsic" average connectivity $q_1, \ldots, q_n \sim \mu$ i.i.d.
- k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ independent of $\{q_i\}$ of (large) sizes $n_1, \ldots, n_k$, with preferential attachment C_{ab} between $\mathcal{C}_a$ and $\mathcal{C}_b$
- edge probability for nodes $i \in \mathcal{C}_{g_i}$:

$$P(i \sim j) = q_i q_j C_{g_i g_j}.$$

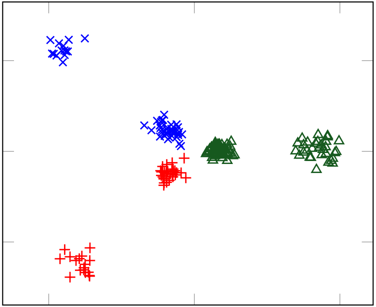
- adjacency matrix A with

$$\boxed{A_{ij} \sim \mathrm{Bernoulli}(q_i q_j C_{g_i g_j})}$$

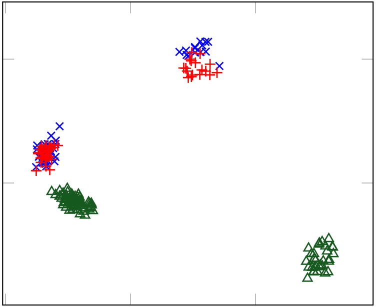
Limitations of Classical Methods

- 3 classes with μ bi-modal ($\mu = \frac{3}{4}\boldsymbol{\delta}_{0.1} + \frac{1}{4}\boldsymbol{\delta}_{0.5}$)

- 3 classes with μ bi-modal ($\mu = \frac{3}{4}\boldsymbol{\delta}_{0.1} + \frac{1}{4}\boldsymbol{\delta}_{0.5}$)



(Modularity $A - \frac{dd^\mathsf{T}}{2m}$)



(Bethe Hessian $D - rA$)

Recall: $P(i \sim j) = q_i q_j C_{g_i g_j}$.

Proposed Regularized Modularity Approach

Recall: $P(i \sim j) = q_i q_j C_{g_i g_j}$.

Dense Regime Assumptions: Non trivial regime when, $\forall a, b$, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}},\ M_{ab} = O(1).$$

Recall: $P(i \sim j) = q_i q_j C_{g_i g_j}$.

Dense Regime Assumptions: Non trivial regime when, $\forall a, b$, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}},\ M_{ab} = O(1).$$

Community information is **weak but highly redundant**

Recall: $P(i \sim j) = q_i q_j C_{g_i g_j}$.

Dense Regime Assumptions: Non trivial regime when, $\forall a, b$, as $n \to \infty$,

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}},\ M_{ab} = O(1).$$

$$\boxed{\text{Community information is } \textbf{weak but highly redundant}}$$

Considered Matrix:

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}.$$

Theorem (Limiting Random Matrix Equivalent)

As $n \to \infty$, $\|L_\alpha - \tilde{L}_\alpha\| \xrightarrow{\text{a.s.}} 0$, where

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T}$$

with $D_q = \operatorname{diag}(\{q_i\})$, X zero-mean random matrix with variance profile,

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D_q^{-\alpha} X 1_n \end{bmatrix}, \quad \textit{rank } k+1$$

$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\mathsf{T}) M (I_k - c 1_k^\mathsf{T}) & -1_k \\ 1_k^\mathsf{T} & 0 \end{bmatrix}$$

and $J = [j_1, \ldots, j_k]$, $j_a = [0, \ldots, 0, 1_{n_a}^\mathsf{T}, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^n$.

Theorem (Limiting Random Matrix Equivalent)

As $n \to \infty$, $\|L_\alpha - \tilde{L}_\alpha\| \xrightarrow{\text{a.s.}} 0$, *where*

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T}$$

with $D_q = \operatorname{diag}(\{q_i\})$, X *zero-mean random matrix with variance profile,*

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D_q^{-\alpha} X 1_n \end{bmatrix}, \quad \textit{rank } k+1$$

$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\mathsf{T}) M (I_k - c 1_k^\mathsf{T}) & -1_k \\ 1_k^\mathsf{T} & 0 \end{bmatrix}$$

and $J = [j_1, \ldots, j_k]$, $j_a = [0, \ldots, 0, 1_{n_a}^\mathsf{T}, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^n$.

Consequences:

- isolated eigenvalues beyond phase transition $\Leftrightarrow \lambda(M) >$ "spectrum edge"

$\boxed{\text{Optimal choice } \alpha_{\text{opt}} \text{ of } \alpha \text{ from study of limiting spectrum.}}$

Theorem (Limiting Random Matrix Equivalent)

As $n \to \infty$, $\|L_\alpha - \tilde{L}_\alpha\| \xrightarrow{\text{a.s.}} 0$, *where*

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[A - \frac{dd^\mathsf{T}}{2m} \right] D^{-\alpha}$$

$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\mathsf{T}$$

with $D_q = \mathrm{diag}(\{q_i\})$, X *zero-mean random matrix with variance profile,*

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & D_q^{-\alpha} X 1_n \end{bmatrix}, \quad \textit{rank } k+1$$

$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\mathsf{T}) M (I_k - c 1_k^\mathsf{T}) & -1_k \\ 1_k^\mathsf{T} & 0 \end{bmatrix}$$

and $J = [j_1, \ldots, j_k]$, $j_a = [0, \ldots, 0, 1_{n_a}^\mathsf{T}, 0, \ldots, 0]^\mathsf{T} \in \mathbb{R}^n$.

Consequences:

- isolated eigenvalues beyond phase transition $\Leftrightarrow$ $\lambda(M) >$ "spectrum edge"

  Optimal choice α_{opt} of α from study of limiting spectrum.

- **eigenvectors correlated to** $D_q^{1-\alpha} J$

  Necessary regularization by $D^{\alpha-1}$.

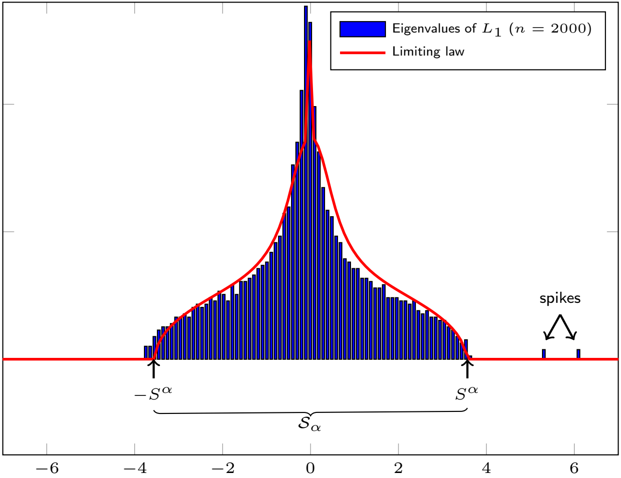


Figure: 3 classes, $c_1 = c_2 = 0.3, c_3 = 0.4$, $\mu = \frac{1}{2}\boldsymbol{\delta}_{0.4} + \frac{1}{2}\boldsymbol{\delta}_{0.9}$, $M = 4\begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}$.

Theorem (Phase Transition)

Isolated eigenvalue $\lambda_i(L_\alpha)$ *if* $|\lambda_i(\bar{M})| > \tau^\alpha$, $\bar{M} = (\mathcal{D}(c) - cc^\mathsf{T})M$, *where*

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{g^\alpha(x)}, \text{ phase transition threshold}$$

with $[S_-^\alpha, S_+^\alpha]$ *limiting eigenvalue support of* L_α *and* $g^\alpha(x)$ *(*$|x| > S_+^\alpha$*) solution of*

$$f^\alpha(x) = \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq)$$
$$g^\alpha(x) = \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq).$$

In this case, $\lambda_i(L_\alpha) \xrightarrow{\text{a.s.}} (g^\alpha)^{-1} \left(-1/\lambda_i(\bar{M})\right)$.

Theorem (Phase Transition)

Isolated eigenvalue $\lambda_i(L_\alpha)$ *if* $|\lambda_i(\bar{M})| > \tau^\alpha$, $\bar{M} = (\mathcal{D}(c) - cc^\mathsf{T})M$, *where*

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{g^\alpha(x)}, \text{ phase transition threshold}$$

with $[S_-^\alpha, S_+^\alpha]$ *limiting eigenvalue support of* L_α *and* $g^\alpha(x)$ ($|x| > S_+^\alpha$) *solution of*

$$f^\alpha(x) = \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq)$$

$$g^\alpha(x) = \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq).$$

In this case, $\lambda_i(L_\alpha) \xrightarrow{\text{a.s.}} (g^\alpha)^{-1}\left(-1/\lambda_i(\bar{M})\right)$.

Clustering possible when $\lambda_i(\bar{M}) > (\min_\alpha \tau_\alpha)$:

- **"Optimal"** $\alpha_{\text{opt}} \equiv \operatorname{argmin}_\alpha \{\tau_\alpha\}$.

Theorem (Phase Transition)

Isolated eigenvalue $\lambda_i(L_\alpha)$ *if* $|\lambda_i(\bar{M})| > \tau^\alpha$, $\bar{M} = (\mathcal{D}(c) - cc^\mathsf{T})M$, *where*

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{g^\alpha(x)}, \text{ phase transition threshold}$$

with $[S_-^\alpha, S_+^\alpha]$ *limiting eigenvalue support of* L_α *and* $g^\alpha(x)$ *(*$|x| > S_+^\alpha$*) solution of*

$$f^\alpha(x) = \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq)$$

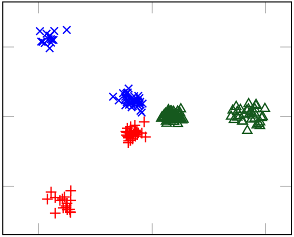
$$g^\alpha(x) = \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} f^\alpha(x) + q^{2-2\alpha} g^\alpha(x)} \mu(dq).$$

In this case, $\lambda_i(L_\alpha) \xrightarrow{\text{a.s.}} (g^\alpha)^{-1}\left(-1/\lambda_i(\bar{M})\right)$.

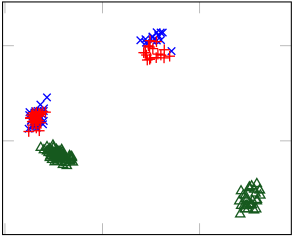
Clustering possible when $\lambda_i(\bar{M}) > (\min_\alpha \tau_\alpha)$:

- **"Optimal"** $\alpha_{\text{opt}} \equiv \operatorname{argmin}_\alpha \{\tau_\alpha\}$.
- From $\hat{q}_i \equiv \frac{d_i}{\sqrt{d^\mathsf{T} 1_n}} \xrightarrow{\text{a.s.}} q_i$, $\mu \simeq \hat{\mu} \equiv \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{\hat{q}_i}$ and thus:

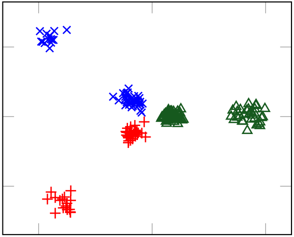
$$\boxed{\text{Consistent estimator } \hat{\alpha}_{\text{opt}} \text{ of } \alpha_{\text{opt}}.}$$



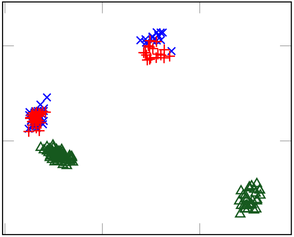
(Modularity $A - \frac{dd^{\mathsf{T}}}{2m}$)



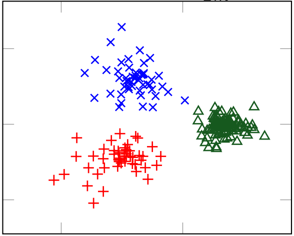
(Bethe Hessian $D - rA$)



(Modularity $A - \frac{dd^{\mathsf{T}}}{2m}$)

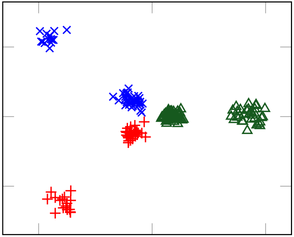


(Bethe Hessian $D - rA$)

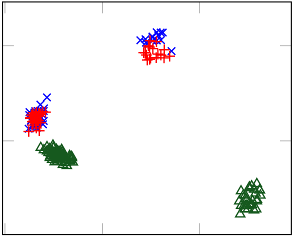


(Proposed, $\alpha = 1$)

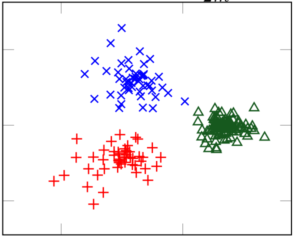
Figure: 3 classes, $\mu = \frac{3}{4}\boldsymbol{\delta}_{0.1} + \frac{1}{4}\boldsymbol{\delta}_{0.5}$, $c_1 = c_2 = \frac{1}{4}$, $c_3 = \frac{1}{2}$, $M = 100 I_3$.



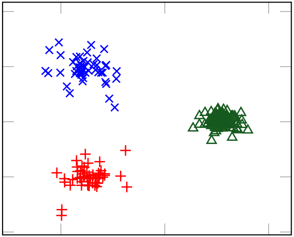
(Modularity $A - \frac{dd^{\mathsf{T}}}{2m}$)



(Bethe Hessian $D - rA$)



(Proposed, $\alpha = 1$)



(Proposed, $\hat{\alpha}_{\text{opt}}$)

Figure: 3 classes, $\mu = \frac{3}{4}\boldsymbol{\delta}_{0.1} + \frac{1}{4}\boldsymbol{\delta}_{0.5}$, $c_1 = c_2 = \frac{1}{4}$, $c_3 = \frac{1}{2}$, $M = 100I_3$.

Simulated Performance Results (2 masses for q_i)

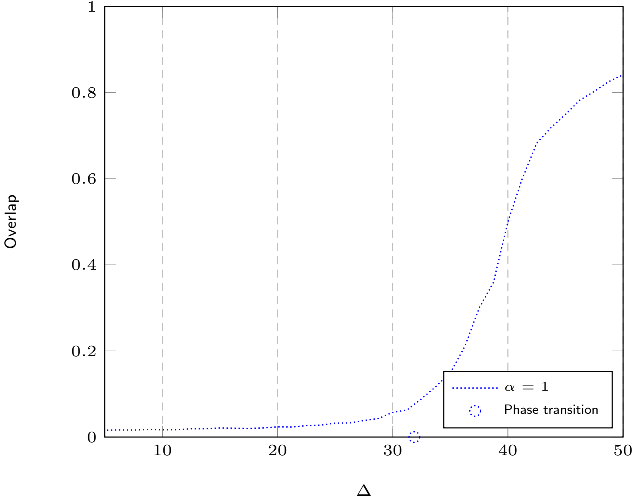


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\rm opt} = 0.07$.

Simulated Performance Results (2 masses for q_i)

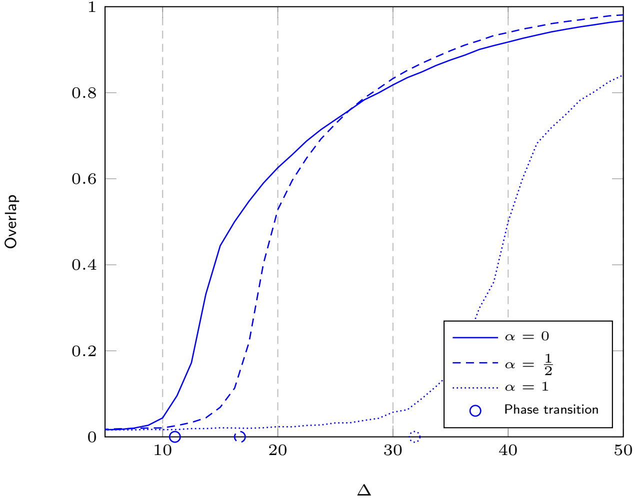


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\rm opt} = 0.07$.

Simulated Performance Results (2 masses for q_i)

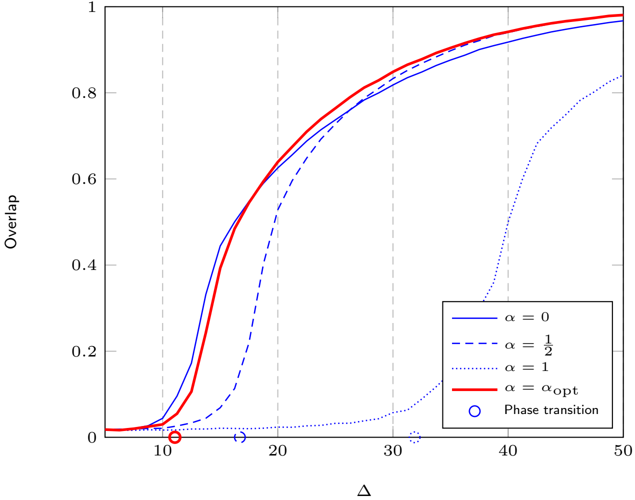


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\mathrm{opt}} = 0.07$.

Simulated Performance Results (2 masses for q_i)

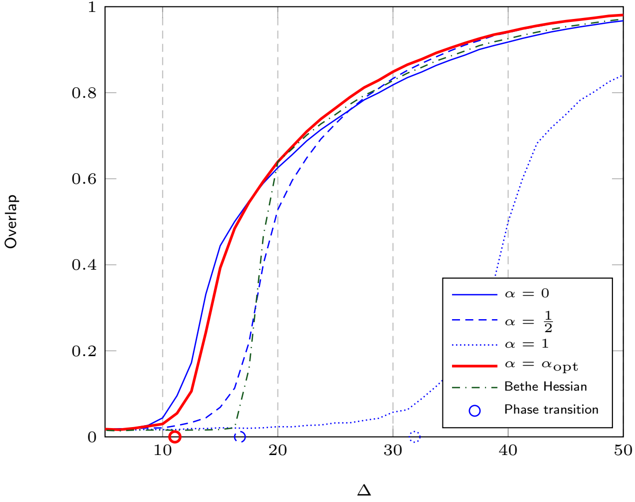


Figure: Overlap performance for $n = 3000$, $K = 3$, $c_i = \frac{1}{3}$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} = 0.5$, $M = \Delta I_3$, for $\Delta \in [5, 50]$. Here $\alpha_{\text{opt}} = 0.07$.

Simulated Performance Results (2 masses for q_i)

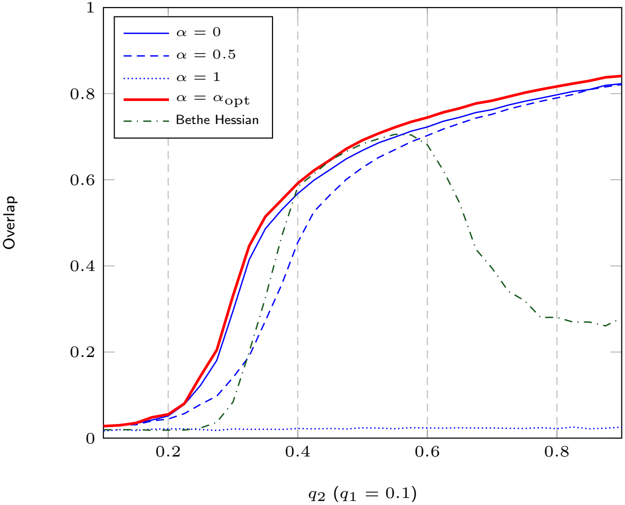
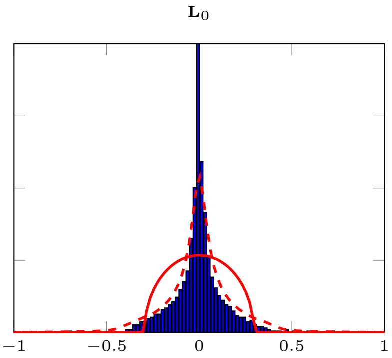
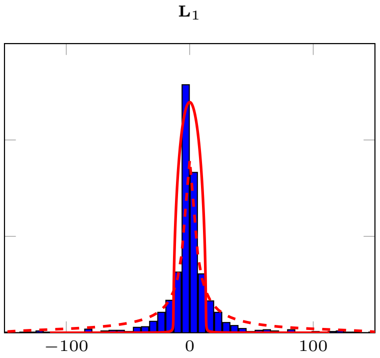


Figure: Overlap performance for $n = 3000$, $K = 3$, $\mu = \frac{3}{4}\delta_{q_{(1)}} + \frac{1}{4}\delta_{q_{(2)}}$ with $q_{(1)} = 0.1$ and $q_{(2)} \in [0.1, 0.9]$, $M = 10(2I_3 - 1_3 1_3^{\mathsf{T}})$, $c_i = \frac{1}{3}$.

Real Graph Example: PolBlogs ($n = 1490$, two classes)



($\lambda_{\max} \simeq 1.75$)



($\lambda_{\max} \simeq 483$)

Algorithms	Overlap	Modularity
α_{opt} ($\simeq 0$)	**0.897**	**0.4246**
$\alpha = 0.5$	0.035	$\simeq 0$
$\alpha = 1$	0.040	$\simeq 0$
BH	0.304	0.2723

Outline

Summary of Results and Perspectives I

Random Neural Networks.

- ✔ Extreme learning machines (one-layer random NN)
- ✔ Linear echo-state networks (ESN)
- ✎ Logistic regression and classification error in extreme learning machines (ELM)
- ✎ Further random feature maps characterization
- ✎ Generalized random NN (multiple layers, multiple activations)
- ✎ Random convolutional networks for image processing
- Non-linear ESN

Deep Neural Networks (DNN).

- ✎ Backpropagation in NN ($\sigma(WX)$ for random X, backprop. on W)
- Statistical physics-inspired approaches (spin-glass models, Hamiltonian-based models)
- Non-linear ESN

DNN performance of physics-realistic models (4th-order Hamiltonian, locality)

References.

- H. W. Lin, M. Tegmark, “Why does deep and cheap learning work so well?”, arXiv:1608.08225v2, 2016.
- C. Williams, “Computation with infinite neural networks”, Neural Computation, 10(5), 1203-1216, 1998.
- Herbert Jaeger. Short term memory in echo state networks. GMD-Forschungszentrum Informationstechnik, 2001.
- Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew, “Extreme learning machine : theory and applications”, Neurocomputing, 70(1) :489501, 2006.
- N. El Karoui, “Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond”, The Annals of Applied Probability, 19(6), 2362-2405, 2009.
- C. Louart, Z. Liao, R. Couillet, ”A Random Matrix Approach to Neural Networks”, (submitted to) Annals of Applied Probability, 2017.
- R. Couillet, G. Wainrib, H. Sevi, H. Tiomoko Ali, “The asymptotic performance of linear echo state neural networks”, Journal of Machine Learning Research, vol. 17, no. 178, pp. 1-35, 2016.
- Choromanska, Anna, et al. ”The Loss Surfaces of Multilayer Networks.” AISTATS. 2015.
- Rahimi, Ali, and Benjamin Recht. ”Random Features for Large-Scale Kernel Machines.” NIPS. Vol. 3. No. 4. 2007.

Kernel methods.

- ✔ Spectral clustering
- ✔ Subspace spectral clustering ($f'(\tau) = 0$)
- ✎ Spectral clustering with outer product kernel $f(x^\mathsf{T} y)$
- ✔ Semi-supervised learning, kernel approaches.
- ✔ Least square support vector machines (LS-SVM).
- ✎ Support vector machines (SVM).
- Kernel matrices based on Kendall τ, Spearman ρ.

Applications.

- ✔ Massive MIMO user subspace clustering (patent proposed)
- Kernel correlation matrices for biostats, heterogeneous datasets.
- Kernel PCA.
- Kendall τ in biostats.

References.

N. El Karoui, "The spectrum of kernel random matrices", The Annals of Statistics, 38(1), 1-50, 2010.

Summary of Results and Perspectives II

R. Couillet, F. Benaych-Georges, “Kernel Spectral Clustering of Large Dimensional Data”, Electronic Journal of Statistics, vol. 10, no. 1, pp. 1393-1454, 2016.

R. Couillet, A. Kammoun, “Random Matrix Improved Subspace Clustering”, Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2016.

Z. Liao, R. Couillet, ”A Large Dimensional Analysis of Least Squares Support Vector Machines”, (submitted to) Journal of Machine Learning Research, 2017.

X. Mai, R. Couillet, “The counterintuitive mechanism of graph-based semi-supervised learning in the big data regime”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’17), New Orleans, USA, 2017.

Community detection.

- ✔ Heterogeneous dense network clustering.
- ✎ Semi-supervised clustering.
- 💡 Sparse network extensions.
- 💡 Beyond community detection (hub detection).

Applications.

- ✔ Improved methods for community detection.
- ✎ Applications to distributed optimization (network diffusion, graph signal processing).

References.

H. Tiomoko Ali, R. Couillet, "Spectral community detection in heterogeneous large networks", (submitted to) Journal of Multivariate Analysis, 2016.

F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, P. Zhang, "Spectral redemption in clustering sparse networks. Proceedings of the National Academy of Sciences", 110(52), 20935-20940, 2013.

C. Bordenave, M. Lelarge, L. Massoulié, "Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs", Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on, pp. 1347-1357, 2015

A. Saade, F. Krzakala, L. Zdeborová, "Spectral clustering of graphs with the Bethe Hessian", In Advances in Neural Information Processing Systems, pp. 406-414, 2014.

Summary of Results and Perspectives I

Robust statistics.

- ✔ Tyler, Maronna (and regularized) estimators
- ✔ Elliptical data setting, deterministic outlier setting
- ✔ Central limit theorem extensions
- Joint mean and covariance robust estimation
- Robust regression (preliminary works exist already using strikingly different approaches)

Applications.

- ✔ Statistical finance (portfolio estimation)
- ✔ Localisation in array processing (robust GMUSIC)
- ✔ Detectors in space time array processing
- Correlation matrices in biostatistics, human science datasets, etc.

References.

R. Couillet, F. Pascal, J. W. Silverstein, "Robust Estimates of Covariance Matrices in the Large Dimensional Regime", IEEE Transactions on Information Theory, vol. 60, no. 11, pp. 7269-7278, 2014.

R. Couillet, F. Pascal, J. W. Silverstein, "The Random Matrix Regime of Maronna's M-estimator with elliptically distributed samples", Elsevier Journal of Multivariate Analysis, vol. 139, pp. 56-78, 2015.

T. Zhang, X. Cheng, A. Singer, "Marchenko-Pastur Law for Tyler's and Maronna's M-estimators", arXiv:1401.3424, 2014.

R. Couillet, M. McKay, "Large Dimensional Analysis and Optimization of Robust Shrinkage Covariance Matrix Estimators", Elsevier Journal of Multivariate Analysis, vol. 131, pp. 99-120, 2014.

D. Morales-Jimenez, R. Couillet, M. McKay, "Large Dimensional Analysis of Robust M-Estimators of Covariance with Outliers", IEEE Transactions on Signal Processing, vol. 63, no. 21, pp. 5784-5797, 2015.

L. Yang, R. Couillet, M. McKay, "A Robust Statistics Approach to Minimum Variance Portfolio Optimization", IEEE Transactions on Signal Processing, vol. 63, no. 24, pp. 6684–6697, 2015.

R. Couillet, "Robust spiked random matrices and a robust G-MUSIC estimator", Elsevier Journal of Multivariate Analysis, vol. 140, pp. 139-161, 2015.

A. Kammoun, R. Couillet, F. Pascal, M.-S. Alouini, "Optimal Design of the Adaptive Normalized Matched Filter Detector", (submitted to) IEEE Transactions on Information Theory, 2016, arXiv Preprint 1504.01252.

R. Couillet, A. Kammoun, F. Pascal, "Second order statistics of robust estimators of scatter. Application to GLRT detection for elliptical signals", Elsevier Journal of Multivariate Analysis, vol. 143, pp. 249-274, 2016.

D. Donoho, A. Montanari, "High dimensional robust m-estimation: Asymptotic variance via approximate message passing", Probability Theory and Related Fields, 1-35, 2013.

N. El Karoui, "Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results." arXiv preprint arXiv:1311.2445, 2013.

Summary of Results and Perspectives I

Other works and ideas.

- ✔ Spike random matrix sparse PCA
- ✎ Non-linear shrinkage methods
- ✎ Sparse kernel PCA
- ✎ Random signal processing on graph methods.
- ✎ Random matrix analysis of diffusion networks performance.

Applications.

- ✔ Spike factor models in portfolio optimization
- ✎ Non-linear shrinkage in portfolio optimization, biostats

References.

R. Couillet, M. McKay, "Optimal block-sparse PCA for high dimensional correlated samples", (submitted to) Journal of Multivariate Analysis, 2016.

J. Bun, J. P. Bouchaud, M. Potters, "On the overlaps between eigenvectors of correlated random matrices", arXiv preprint arXiv:1603.04364 (2016).

Ledoit, O. and Wolf, M., "Nonlinear shrinkage estimation of large-dimensional covariance matrices", 2011

Thank you.