

# A Random Matrix Approach to Machine Learning

(XII Brunel – Bielefeld Workshop on RMT)

Romain COUILLET

CentraleSupélec, France

December, 2016



CentraleSupélec

Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

Kernel Spectral Clustering

Kernel Spectral Clustering: Subspace Clustering

Semi-supervised Learning

Support Vector Machines

Neural Networks: Extreme Learning Machines

Random Matrices and Robust Estimation

Perspectives

## Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

Kernel Spectral Clustering

Kernel Spectral Clustering: Subspace Clustering

Semi-supervised Learning

Support Vector Machines

Neural Networks: Extreme Learning Machines

Random Matrices and Robust Estimation

Perspectives

## Reminder on Spectral Clustering Methods

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

1. extraction of eigenvectors  $U = [u_1, \dots, u_\ell]$  with “dominant” eigenvalues

## Reminder on Spectral Clustering Methods

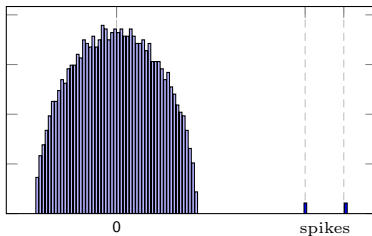
**Context:** Two-step classification of  $n$  objects based on similarity  $A \in \mathbb{R}^{n \times n}$ :

1. extraction of eigenvectors  $U = [u_1, \dots, u_\ell]$  with “dominant” eigenvalues
2. classification of vectors  $U_{\cdot,1}, \dots, U_{\cdot,n} \in \mathbb{R}^\ell$  using k-means/EM.

## Reminder on Spectral Clustering Methods

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

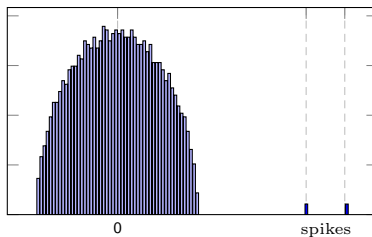
1. extraction of eigenvectors  $U = [u_1, \dots, u_\ell]$  with “dominant” eigenvalues
2. classification of vectors  $U_{\cdot,1}, \dots, U_{\cdot,n} \in \mathbb{R}^\ell$  using k-means/EM.



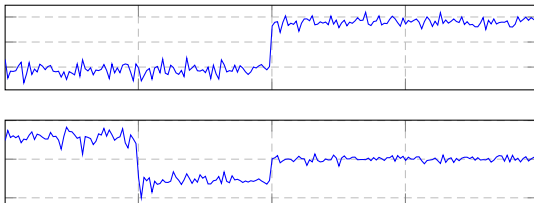
## Reminder on Spectral Clustering Methods

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

1. extraction of eigenvectors  $U = [u_1, \dots, u_\ell]$  with “dominant” eigenvalues
2. classification of vectors  $U_{\cdot,1}, \dots, U_{\cdot,n} \in \mathbb{R}^\ell$  using k-means/EM.



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



## Reminder on Spectral Clustering Methods

Eigenv. 1



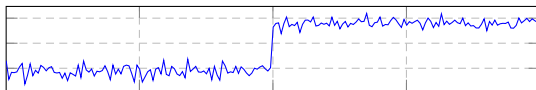
Eigenv. 2





## Reminder on Spectral Clustering Methods

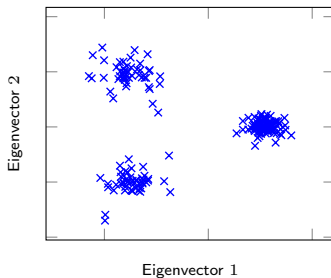
Eigenv. 1



Eigenv. 2

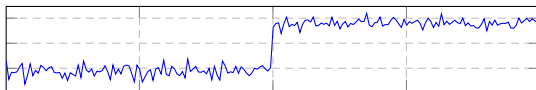


↓  $\ell$ -dimensional representation ↓  
(shuffling no longer matters!)



## Reminder on Spectral Clustering Methods

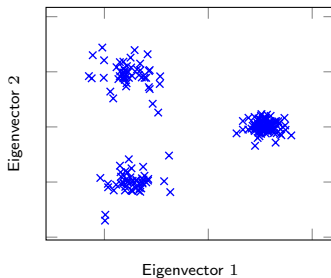
Eigenv. 1



Eigenv. 2



↓  $\ell$ -dimensional representation ↓  
(shuffling no longer matters!)



↓  
**EM or k-means clustering.**

# The Random Matrix Approach

## A two-step method:

1. If  $A_n$  is not a “standard” random matrix, retrieve  $\tilde{A}_n$  such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as  $n \rightarrow \infty$ .

# The Random Matrix Approach

## A two-step method:

1. If  $A_n$  is not a “standard” random matrix, retrieve  $\tilde{A}_n$  such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as  $n \rightarrow \infty$ .

$\Rightarrow$  Transfers crucial properties from  $A_n$  to  $\tilde{A}_n$ :

# The Random Matrix Approach

## A two-step method:

1. If  $A_n$  is not a “standard” random matrix, retrieve  $\tilde{A}_n$  such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as  $n \rightarrow \infty$ .

⇒ Transfers crucial properties from  $A_n$  to  $\tilde{A}_n$ :

- ▶ limiting eigenvalue distribution

# The Random Matrix Approach

## A two-step method:

1. If  $A_n$  is not a “standard” random matrix, retrieve  $\tilde{A}_n$  such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as  $n \rightarrow \infty$ .

⇒ Transfers crucial properties from  $A_n$  to  $\tilde{A}_n$ :

- ▶ limiting eigenvalue distribution
- ▶ spikes

# The Random Matrix Approach

## A two-step method:

1. If  $A_n$  is not a “standard” random matrix, retrieve  $\tilde{A}_n$  such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as  $n \rightarrow \infty$ .

⇒ Transfers crucial properties from  $A_n$  to  $\tilde{A}_n$ :

- ▶ limiting eigenvalue distribution
- ▶ spikes
- ▶ eigenvectors of isolated eigenvalues.

# The Random Matrix Approach

## A two-step method:

1. If  $A_n$  is not a “standard” random matrix, retrieve  $\tilde{A}_n$  such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as  $n \rightarrow \infty$ .

⇒ Transfers crucial properties from  $A_n$  to  $\tilde{A}_n$ :

- ▶ limiting eigenvalue distribution
- ▶ spikes
- ▶ eigenvectors of isolated eigenvalues.

2. From  $\tilde{A}$ , perform spiked model analysis:



# The Random Matrix Approach

## A two-step method:

1. If  $A_n$  is not a “standard” random matrix, retrieve  $\tilde{A}_n$  such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as  $n \rightarrow \infty$ .

⇒ Transfers crucial properties from  $A_n$  to  $\tilde{A}_n$ :

- ▶ limiting eigenvalue distribution
  - ▶ spikes
  - ▶ eigenvectors of isolated eigenvalues.
2. From  $\tilde{A}$ , perform spiked model analysis:
    - ▶ exhibit phase transition phenomenon

# The Random Matrix Approach

## A two-step method:

1. If  $A_n$  is not a “standard” random matrix, retrieve  $\tilde{A}_n$  such that

$$\left\| A_n - \tilde{A}_n \right\| \xrightarrow{\text{a.s.}} 0$$

in operator norm as  $n \rightarrow \infty$ .

⇒ Transfers crucial properties from  $A_n$  to  $\tilde{A}_n$ :

- ▶ limiting eigenvalue distribution
  - ▶ spikes
  - ▶ eigenvectors of isolated eigenvalues.
2. From  $\tilde{A}$ , perform spiked model analysis:
    - ▶ exhibit phase transition phenomenon
    - ▶ “read” the content of isolated eigenvectors of  $\tilde{A}$ .

# The Random Matrix Approach

## The Spike Analysis:

For “noisy plateaus”-looking isolated eigenvectors  $u_1, \dots, u_\ell$  of  $\tilde{A}$ , write

$$u_i = \sum_{a=1}^k \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

with  $j_a \in \mathbb{R}^n$  canonical vector of class  $\mathcal{C}_a$ ,  $w_i^a$  noise orthogonal to  $j_a$ ,

# The Random Matrix Approach

## The Spike Analysis:

For “noisy plateaus”-looking isolated eigenvectors  $u_1, \dots, u_\ell$  of  $\tilde{A}$ , write

$$u_i = \sum_{a=1}^k \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

with  $j_a \in \mathbb{R}^n$  canonical vector of class  $\mathcal{C}_a$ ,  $w_i^a$  noise orthogonal to  $j_a$ , and evaluate

$$\alpha_i^a = \frac{1}{\sqrt{n_a}} u_i^\top j_a$$
$$(\sigma_i^a)^2 = \left\| u_i - \alpha_i^a \frac{j_a}{\sqrt{n_a}} \right\|^2.$$

# The Random Matrix Approach

## The Spike Analysis:

For “noisy plateaus”-looking isolated eigenvectors  $u_1, \dots, u_\ell$  of  $\tilde{A}$ , write

$$u_i = \sum_{a=1}^k \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

with  $j_a \in \mathbb{R}^n$  canonical vector of class  $\mathcal{C}_a$ ,  $w_i^a$  noise orthogonal to  $j_a$ , and evaluate

$$\alpha_i^a = \frac{1}{\sqrt{n_a}} u_i^\top j_a$$
$$(\sigma_i^a)^2 = \left\| u_i - \alpha_i^a \frac{j_a}{\sqrt{n_a}} \right\|^2.$$

$\Rightarrow$  Can be done using complex analysis calculus, e.g.

$$\begin{aligned} (\alpha_i^a)^2 &= \frac{1}{n_a} j_a^\top u_i u_i^\top j_a \\ &= \frac{1}{2\pi i} \oint_{\gamma_a} \frac{1}{n_a} j_a^\top (\tilde{A} - zI_n)^{-1} j_a dz. \end{aligned}$$

Spectral Clustering Methods and Random Matrices

**Community Detection on Graphs**

Kernel Spectral Clustering

Kernel Spectral Clustering: Subspace Clustering

Semi-supervised Learning

Support Vector Machines

Neural Networks: Extreme Learning Machines

Random Matrices and Robust Estimation

Perspectives

## System Setting

Assume  $n$ -node,  $m$ -edges **undirected** graph  $G$ , with

- ▶ “intrinsic” average connectivity  $q_1, \dots, q_n \sim \mu$  i.i.d.

## System Setting

Assume  $n$ -node,  $m$ -edges **undirected** graph  $G$ , with

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



## System Setting

Assume  $n$ -node,  $m$ -edges **undirected** graph  $G$ , with

- ▶ “intrinsic” average connectivity  $q_1, \dots, q_n \sim \mu$  i.i.d.
- ▶  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$  independent of  $\{q_i\}$  of (large) sizes  $n_1, \dots, n_k$ , with preferential attachment  $C_{ab}$  between  $\mathcal{C}_a$  and  $\mathcal{C}_b$
- ▶ induces edge probability for node  $i \in \mathcal{C}_a$ ,  $j \in \mathcal{C}_b$ ,

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

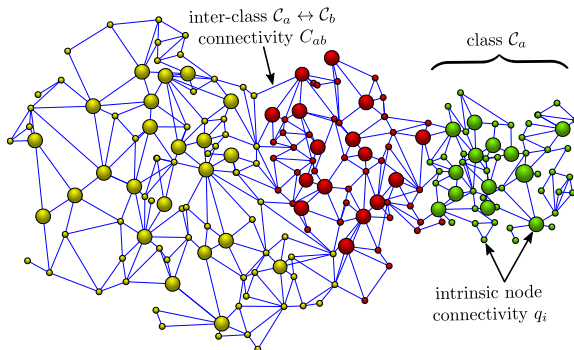
# System Setting

Assume  $n$ -node,  $m$ -edges **undirected** graph  $G$ , with

- ▶ “intrinsic” average connectivity  $q_1, \dots, q_n \sim \mu$  i.i.d.
- ▶  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$  independent of  $\{q_i\}$  of (large) sizes  $n_1, \dots, n_k$ , with preferential attachment  $C_{ab}$  between  $\mathcal{C}_a$  and  $\mathcal{C}_b$
- ▶ induces edge probability for node  $i \in \mathcal{C}_a$ ,  $j \in \mathcal{C}_b$ ,

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

- ▶ adjacency matrix  $A$  with  $A_{ij} \sim \text{Bernoulli}(q_i q_j C_{ab})$ .



## Study of spectral methods:

- ▶ standard methods based on **adjacency**  $A$ , **modularity**  $A - \frac{dd^T}{2m}$ , **normalized adjacency**  $D^{-1}AD^{-1}$ , etc. (adapted to **dense nets**)
- ▶ refined methods based on **Bethe Hessian**  $(r^2 - 1)I_n - rA + D$  (adapted to **sparse nets!**)

## Study of spectral methods:

- ▶ standard methods based on adjacency  $A$ , modularity  $A - \frac{dd^T}{2m}$ , normalized adjacency  $D^{-1}AD^{-1}$ , etc. (adapted to dense nets)
- ▶ refined methods based on Bethe Hessian  $(r^2 - 1)I_n - rA + D$  (adapted to sparse nets!)

## Improvement to realistic graphs:

- ▶ observation of failure of standard methods above
- ▶ improvement by new methods.

## Limitations of Adjacency/Modularity Approach

**Scenario:** 3 classes with  $\mu$  bi-modal (e.g.,  $\mu = \frac{3}{4}\delta_{0.1} + \frac{1}{4}\delta_{0.5}$ )

→ Leading eigenvectors of  $A$  (or modularity  $A - \frac{dd^T}{2m}$ ) **biased by  $q_i$  distribution.**

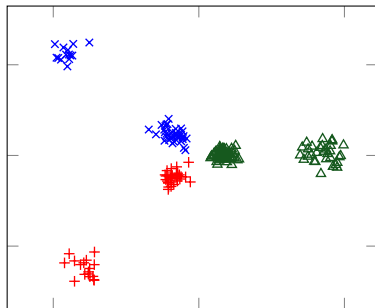
→ Similar behavior for Bethe Hessian.

## Limitations of Adjacency/Modularity Approach

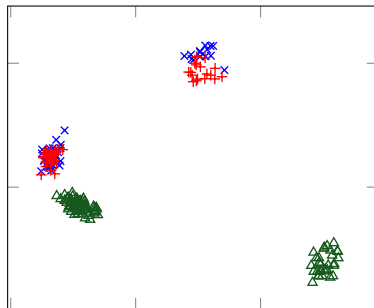
**Scenario:** 3 classes with  $\mu$  bi-modal (e.g.,  $\mu = \frac{3}{4}\delta_{0.1} + \frac{1}{4}\delta_{0.5}$ )

→ Leading eigenvectors of  $A$  (or modularity  $A - \frac{dd^T}{2m}$ ) **biased by  $q_i$  distribution.**

→ Similar behavior for Bethe Hessian.



(Modularity)



(Bethe Hessian)

# Regularized Modularity Approach

**Connectivity Model:**  $P(i \sim j) = q_i q_j C_{ab}$  for  $i \in \mathcal{C}_a, j \in \mathcal{C}_b$ .

**Dense Regime Assumptions:** **Non trivial regime** when, as  $n \rightarrow \infty$ ,

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

# Regularized Modularity Approach

**Connectivity Model:**  $P(i \sim j) = q_i q_j C_{ab}$  for  $i \in \mathcal{C}_a, j \in \mathcal{C}_b$ .

**Dense Regime Assumptions:** **Non trivial regime** when, as  $n \rightarrow \infty$ ,

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

$\Rightarrow$  Community information is **weak but highly REDUNDANT!**



# Regularized Modularity Approach

**Connectivity Model:**  $P(i \sim j) = q_i q_j C_{ab}$  for  $i \in \mathcal{C}_a, j \in \mathcal{C}_b$ .

**Dense Regime Assumptions:** **Non trivial regime** when, as  $n \rightarrow \infty$ ,

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

$\Rightarrow$  Community information is **weak but highly REDUNDANT!**

**Considered Matrix:**

For  $\alpha \in [0, 1]$ , (and with  $D = \text{diag}(A1_n) = \text{diag}(d)$  the degree matrix)

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

# Regularized Modularity Approach

**Connectivity Model:**  $P(i \sim j) = q_i q_j C_{ab}$  for  $i \in \mathcal{C}_a, j \in \mathcal{C}_b$ .

**Dense Regime Assumptions:** **Non trivial regime** when, as  $n \rightarrow \infty$ ,

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

$\Rightarrow$  Community information is **weak but highly REDUNDANT!**

**Considered Matrix:**

For  $\alpha \in [0, 1]$ , (and with  $D = \text{diag}(A1_n) = \text{diag}(d)$  the degree matrix)

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

**Our results in a nutshell:**

- ▶ we find optimal  $\alpha_{\text{opt}}$  having **best phase transition**.

# Regularized Modularity Approach

**Connectivity Model:**  $P(i \sim j) = q_i q_j C_{ab}$  for  $i \in \mathcal{C}_a, j \in \mathcal{C}_b$ .

**Dense Regime Assumptions:** **Non trivial regime** when, as  $n \rightarrow \infty$ ,

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

$\Rightarrow$  Community information is **weak but highly REDUNDANT!**

**Considered Matrix:**

For  $\alpha \in [0, 1]$ , (and with  $D = \text{diag}(A1_n) = \text{diag}(d)$  the degree matrix)

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

**Our results in a nutshell:**

- ▶ we find optimal  $\alpha_{\text{opt}}$  having **best phase transition**.
- ▶ we find **consistent estimator**  $\hat{\alpha}_{\text{opt}}$  from  $A$  alone.

# Regularized Modularity Approach

**Connectivity Model:**  $P(i \sim j) = q_i q_j C_{ab}$  for  $i \in \mathcal{C}_a, j \in \mathcal{C}_b$ .

**Dense Regime Assumptions:** **Non trivial regime** when, as  $n \rightarrow \infty$ ,

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

$\Rightarrow$  Community information is **weak but highly REDUNDANT!**

**Considered Matrix:**

For  $\alpha \in [0, 1]$ , (and with  $D = \text{diag}(A \mathbf{1}_n) = \text{diag}(d)$  the degree matrix)

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

**Our results in a nutshell:**

- ▶ we find optimal  $\alpha_{\text{opt}}$  having **best phase transition**.
- ▶ we find **consistent estimator**  $\hat{\alpha}_{\text{opt}}$  from  $A$  alone.
- ▶ we claim **optimal eigenvector regularization**  $D^{\alpha-1}u$ ,  $u$  eigenvector of  $L_\alpha$ .  
 $\Rightarrow$  **Never proposed before!**

# Asymptotic Equivalence

## Theorem (Limiting Random Matrix Equivalent)

For each  $\alpha \in [0, 1]$ , as  $n \rightarrow \infty$ ,  $\|L_\alpha - \tilde{L}_\alpha\| \rightarrow 0$  almost surely, where

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[ A - \frac{dd^\top}{d^\top 1_n} \right] D^{-\alpha}$$
$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\top$$

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

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & \frac{1}{nm_\mu} D_q^{-\alpha} X 1_n \end{bmatrix}, \text{ rank } k+1$$
$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\top) M (I_k - c 1_k^\top) & -1_k \\ 1_k^\top & 0 \end{bmatrix}$$

and  $J = [j_1, \dots, j_k]$ ,  $j_a = [0, \dots, 0, 1_{n_a}^\top, 0, \dots, 0]^\top \in \mathbb{R}^n$  canonical vector of class  $\mathcal{C}_a$ .

# Asymptotic Equivalence

## Theorem (Limiting Random Matrix Equivalent)

For each  $\alpha \in [0, 1]$ , as  $n \rightarrow \infty$ ,  $\|L_\alpha - \tilde{L}_\alpha\| \rightarrow 0$  almost surely, where

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[ A - \frac{dd^\top}{d^\top 1_n} \right] D^{-\alpha}$$
$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\top$$

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

$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & \frac{1}{nm_\mu} D_q^{-\alpha} X 1_n \end{bmatrix}, \text{ rank } k+1$$
$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\top) M (I_k - c 1_k^\top) & -1_k \\ 1_k^\top & 0 \end{bmatrix}$$

and  $J = [j_1, \dots, j_k]$ ,  $j_a = [0, \dots, 0, 1_{n_a}^\top, 0, \dots, 0]^\top \in \mathbb{R}^n$  canonical vector of class  $\mathcal{C}_a$ .

### Consequences:

- ▶ isolated eigenvalues beyond phase transition  $\leftrightarrow \lambda(M) > \text{"spectrum edge"}$   
 $\Rightarrow$  optimal choice  $\alpha_{\text{opt}}$  of  $\alpha$  from study of noise spectrum.

# Asymptotic Equivalence

## Theorem (Limiting Random Matrix Equivalent)

For each  $\alpha \in [0, 1]$ , as  $n \rightarrow \infty$ ,  $\|L_\alpha - \tilde{L}_\alpha\| \rightarrow 0$  almost surely, where

$$L_\alpha = (2m)^\alpha \frac{1}{\sqrt{n}} D^{-\alpha} \left[ A - \frac{dd^\top}{d^\top 1_n} \right] D^{-\alpha}$$
$$\tilde{L}_\alpha = \frac{1}{\sqrt{n}} D_q^{-\alpha} X D_q^{-\alpha} + U \Lambda U^\top$$

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

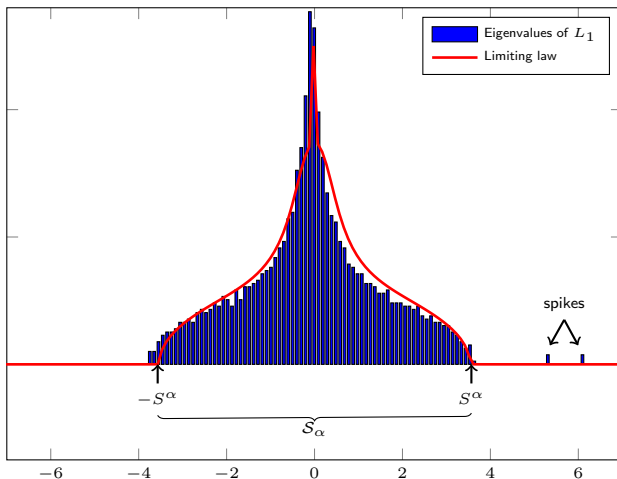
$$U = \begin{bmatrix} D_q^{1-\alpha} \frac{J}{\sqrt{n}} & \frac{1}{nm_\mu} D_q^{-\alpha} X 1_n \end{bmatrix}, \text{ rank } k+1$$
$$\Lambda = \begin{bmatrix} (I_k - 1_k c^\top) M (I_k - c 1_k^\top) & -1_k \\ 1_k^\top & 0 \end{bmatrix}$$

and  $J = [j_1, \dots, j_k]$ ,  $j_a = [0, \dots, 0, 1_{n_a}^\top, 0, \dots, 0]^\top \in \mathbb{R}^n$  canonical vector of class  $\mathcal{C}_a$ .

### Consequences:

- ▶ isolated eigenvalues beyond **phase transition**  $\leftrightarrow \lambda(M) > \text{"spectrum edge"}$   
 $\Rightarrow$  **optimal choice**  $\alpha_{\text{opt}}$  of  $\alpha$  from study of noise spectrum.
- ▶ **eigenvectors correlated to**  $D_q^{1-\alpha} J$   
 $\Rightarrow$  **Natural regularization by**  $D^{\alpha-1} J!$

# Eigenvalue Spectrum



**Figure:** Eigenvalues of  $L_1$ ,  $K = 3$ ,  $n = 2000$ ,  $c_1 = 0.3$ ,  $c_2 = 0.3$ ,  $c_3 = 0.4$ ,  $\mu = \frac{1}{2}\delta_{q_1} + \frac{1}{2}\delta_{q_2}$ ,  $q_1 = 0.4$ ,  $q_2 = 0.9$ ,  $M$  defined by  $M_{ii} = 12$ ,  $M_{ij} = -4$ ,  $i \neq j$ .



## Theorem (Phase Transition)

For  $\alpha \in [0, 1]$ , *isolated eigenvalue*  $\lambda_i(L_\alpha)$  if  $|\lambda_i(\bar{M})| > \tau^\alpha$ ,  $\bar{M} = (\mathcal{D}(c) - cc^\top)M$ ,

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{e_2^\alpha(x)}, \text{ *phase transition threshold*}$$

with  $[S_-^\alpha, S_+^\alpha]$  limiting eigenvalue support of  $m_\mu^{2\alpha} L_\alpha$  and  $e_2^\alpha(x)$  ( $|x| > S_+^\alpha$ ) solution of

$$\begin{aligned} e_1^\alpha(x) &= \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq) \\ e_2^\alpha(x) &= \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq). \end{aligned}$$

In this case,  $-\frac{1}{e_2^\alpha(\lambda_i(m_\mu^{2\alpha} L_\alpha))} = \lambda_i(\bar{M})$ .

## Theorem (Phase Transition)

For  $\alpha \in [0, 1]$ , *isolated eigenvalue*  $\lambda_i(L_\alpha)$  if  $|\lambda_i(\bar{M})| > \tau^\alpha$ ,  $\bar{M} = (\mathcal{D}(c) - cc^\top)M$ ,

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{e_2^\alpha(x)}, \text{ phase transition threshold}$$

with  $[S_-^\alpha, S_+^\alpha]$  limiting eigenvalue support of  $m_\mu^{2\alpha} L_\alpha$  and  $e_2^\alpha(x)$  ( $|x| > S_+^\alpha$ ) solution of

$$\begin{aligned} e_1^\alpha(x) &= \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq) \\ e_2^\alpha(x) &= \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq). \end{aligned}$$

In this case,  $-\frac{1}{e_2^\alpha(\lambda_i(m_\mu^{2\alpha} L_\alpha))} = \lambda_i(\bar{M})$ .

**Clustering still possible** when  $\lambda_i(\bar{M}) = (\min_\alpha \tau_\alpha) + \varepsilon$ .

- “Optimal”  $\alpha = \alpha_{\text{opt}}$ :

$$\alpha_{\text{opt}} = \operatorname{argmin}_{\alpha \in [0,1]} \{\tau_\alpha\}.$$

# Phase Transition

## Theorem (Phase Transition)

For  $\alpha \in [0, 1]$ , *isolated eigenvalue*  $\lambda_i(L_\alpha)$  if  $|\lambda_i(\bar{M})| > \tau^\alpha$ ,  $\bar{M} = (\mathcal{D}(c) - cc^\top)M$ ,

$$\tau^\alpha = \lim_{x \downarrow S_+^\alpha} -\frac{1}{e_2^\alpha(x)}, \text{ *phase transition threshold*}$$

with  $[S_-^\alpha, S_+^\alpha]$  limiting eigenvalue support of  $m_\mu^{2\alpha} L_\alpha$  and  $e_2^\alpha(x)$  ( $|x| > S_+^\alpha$ ) solution of

$$\begin{aligned} e_1^\alpha(x) &= \int \frac{q^{1-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq) \\ e_2^\alpha(x) &= \int \frac{q^{2-2\alpha}}{-x - q^{1-2\alpha} e_1^\alpha(x) + q^{2-2\alpha} e_2^\alpha(x)} \mu(dq). \end{aligned}$$

In this case,  $-\frac{1}{e_2^\alpha(\lambda_i(m_\mu^{2\alpha} L_\alpha))} = \lambda_i(\bar{M})$ .

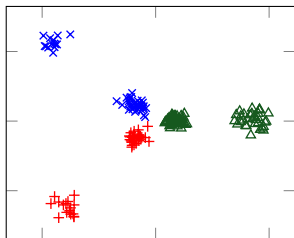
**Clustering still possible** when  $\lambda_i(\bar{M}) = (\min_\alpha \tau_\alpha) + \varepsilon$ .

- “Optimal”  $\alpha = \alpha_{\text{opt}}$ :

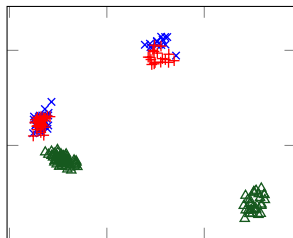
$$\alpha_{\text{opt}} = \operatorname{argmin}_{\alpha \in [0,1]} \{\tau_\alpha\}.$$

- From  $\max_i \left| \frac{d_i}{\sqrt{d^\top 1_n}} - q_i \right| \xrightarrow{\text{a.s.}} 0$ , we obtain **consistent estimator**  $\hat{\alpha}_{\text{opt}}$  of  $\alpha_{\text{opt}}$ .

## Simulated Performance Results (2 masses of $q_i$ )

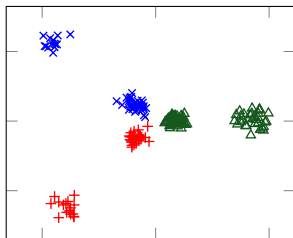


(Modularity)

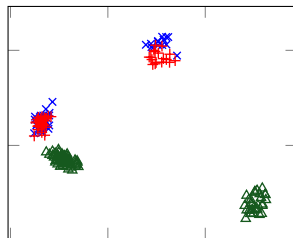


(Bethe Hessian)

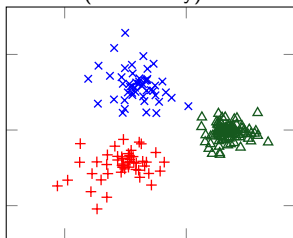
## Simulated Performance Results (2 masses of $q_i$ )



(Modularity)



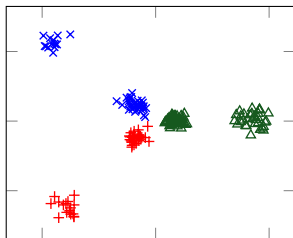
(Bethe Hessian)



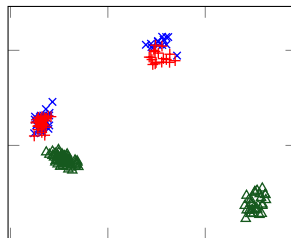
(Algo with  $\alpha = 1$ )

**Figure:** Two dominant eigenvectors (x-y axes) for  $n = 2000$ ,  $K = 3$ ,  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ ,  $q_1 = 0.1$ ,  $q_2 = 0.5$ ,  $c_1 = c_2 = \frac{1}{4}$ ,  $c_3 = \frac{1}{2}$ ,  $M = 100I_3$ .

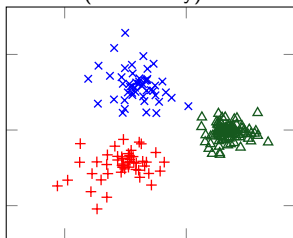
## Simulated Performance Results (2 masses of $q_i$ )



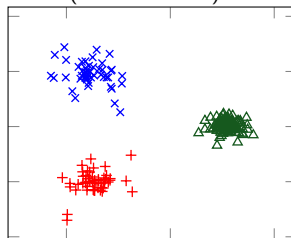
(Modularity)



(Bethe Hessian)



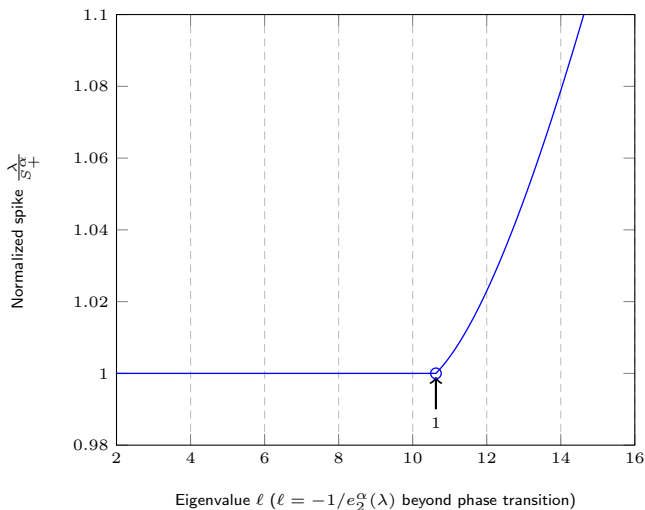
(Algo with  $\alpha = 1$ )



(Algo with  $\alpha_{\text{opt}}$ )

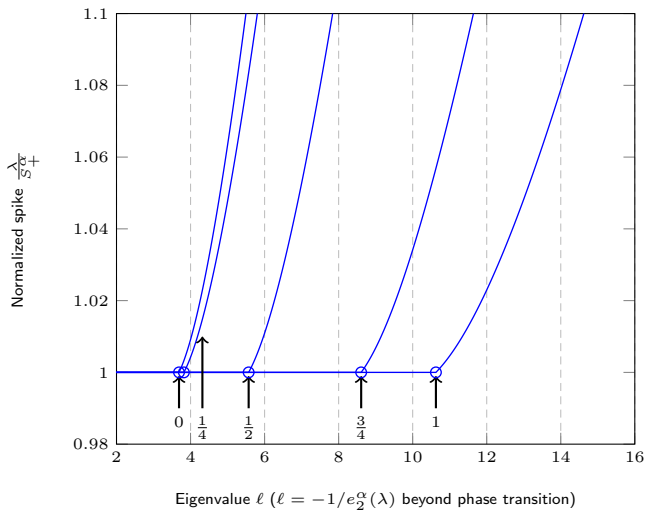
**Figure:** Two dominant eigenvectors (x-y axes) for  $n = 2000$ ,  $K = 3$ ,  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$ ,  $q_1 = 0.1$ ,  $q_2 = 0.5$ ,  $c_1 = c_2 = \frac{1}{4}$ ,  $c_3 = \frac{1}{2}$ ,  $M = 100I_3$ .

## Simulated Performance Results (2 masses for $q_i$ )



**Figure:** Largest eigenvalue  $\lambda$  of  $L_\alpha$  as a function of the largest eigenvalue  $\ell$  of  $(\mathcal{D}(c) - cc^\top)M$ , for  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$  with  $q_1 = 0.1$  and  $q_2 = 0.5$ , for  $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\text{opt}}\}$  (indicated below the graph). Here,  $\alpha_{\text{opt}} = 0.07$ . Circles indicate phase transition. Beyond phase transition,  $\ell = -1/e_2^\alpha(\lambda)$ .

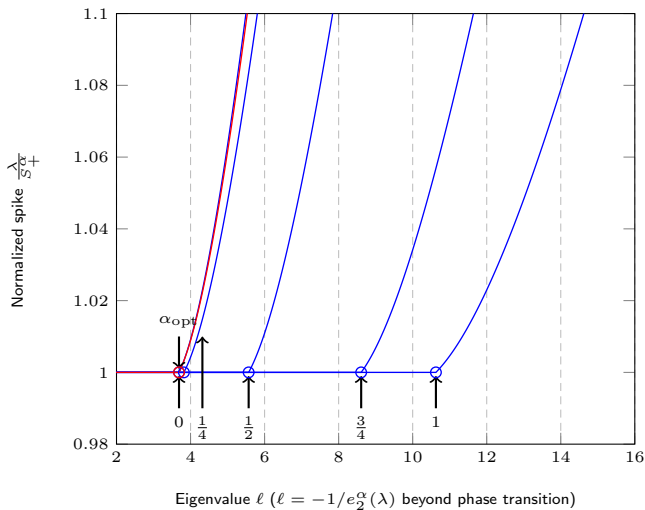
## Simulated Performance Results (2 masses for $q_i$ )



**Figure:** Largest eigenvalue  $\lambda$  of  $L_\alpha$  as a function of the largest eigenvalue  $\ell$  of  $(\mathcal{D}(c) - cc^\top)M$ , for  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$  with  $q_1 = 0.1$  and  $q_2 = 0.5$ , for  $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\text{opt}}\}$  (indicated below the graph). Here,  $\alpha_{\text{opt}} = 0.07$ . Circles indicate phase transition. Beyond phase transition,  $\ell = -1/e_2^\alpha(\lambda)$ .

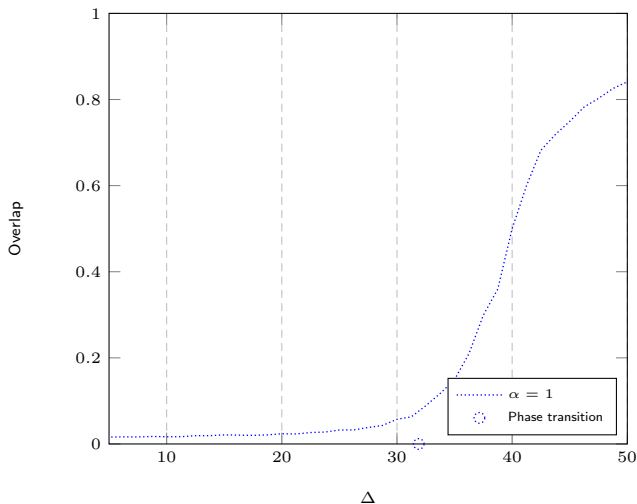


## Simulated Performance Results (2 masses for $q_i$ )



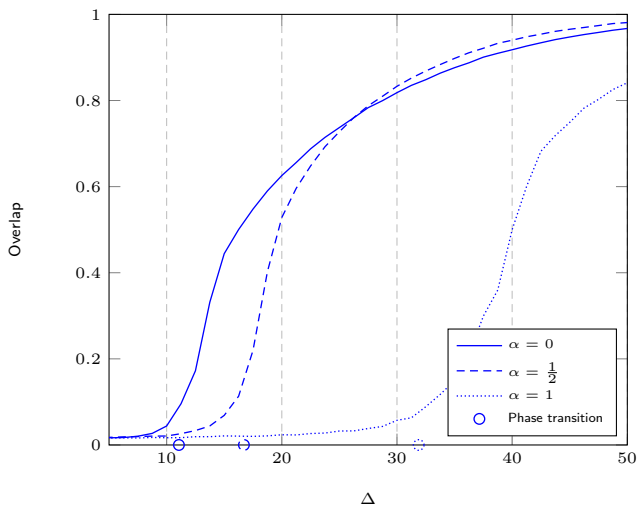
**Figure:** Largest eigenvalue  $\lambda$  of  $L_{\alpha}$  as a function of the largest eigenvalue  $\ell$  of  $(\mathcal{D}(c) - cc^{\top})M$ , for  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$  with  $q_1 = 0.1$  and  $q_2 = 0.5$ , for  $\alpha \in \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1, \alpha_{\text{opt}}\}$  (indicated below the graph). Here,  $\alpha_{\text{opt}} = 0.07$ . Circles indicate phase transition. Beyond phase transition,  $\ell = -1/e_2^{\alpha}(\lambda)$ .

## Simulated Performance Results (2 masses for $q_i$ )



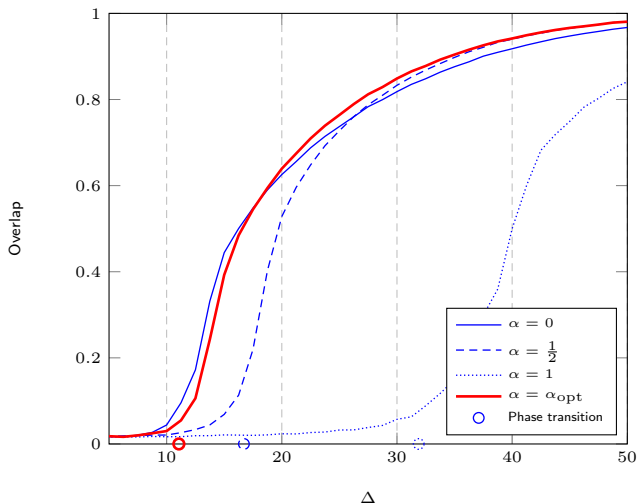
**Figure:** Overlap performance for  $n = 3000$ ,  $K = 3$ ,  $c_i = \frac{1}{3}$ ,  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$  with  $q_1 = 0.1$  and  $q_2 = 0.5$ ,  $M = \Delta I_3$ , for  $\Delta \in [5, 50]$ . Here  $\alpha_{\text{opt}} = 0.07$ .

## Simulated Performance Results (2 masses for $q_i$ )



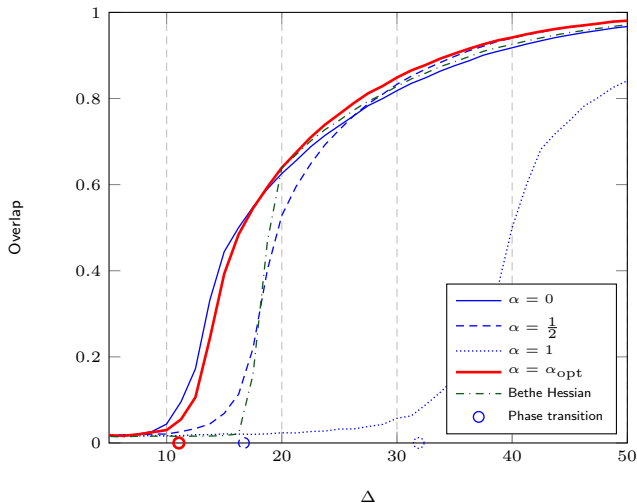
**Figure:** Overlap performance for  $n = 3000$ ,  $K = 3$ ,  $c_i = \frac{1}{3}$ ,  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$  with  $q_1 = 0.1$  and  $q_2 = 0.5$ ,  $M = \Delta I_3$ , for  $\Delta \in [5, 50]$ . Here  $\alpha_{\text{opt}} = 0.07$ .

## Simulated Performance Results (2 masses for $q_i$ )



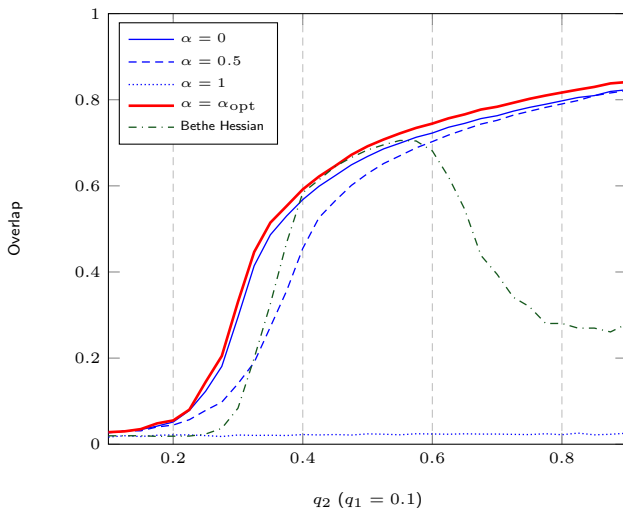
**Figure:** Overlap performance for  $n = 3000$ ,  $K = 3$ ,  $c_i = \frac{1}{3}$ ,  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$  with  $q_1 = 0.1$  and  $q_2 = 0.5$ ,  $M = \Delta I_3$ , for  $\Delta \in [5, 50]$ . Here  $\alpha_{\text{opt}} = 0.07$ .

## Simulated Performance Results (2 masses for $q_i$ )



**Figure:** Overlap performance for  $n = 3000$ ,  $K = 3$ ,  $c_i = \frac{1}{3}$ ,  $\mu = \frac{3}{4}\delta_{q_1} + \frac{1}{4}\delta_{q_2}$  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  $q_2 \in [0.1, 0.9]$ ,  $M = 10(2I_3 - 1_3 1_3^T)$ ,  $c_i = \frac{1}{3}$ .

**Analysis of eigenvectors** reveals:

- ▶ eigenvectors are “noisy staircase vectors”

**Analysis of eigenvectors** reveals:

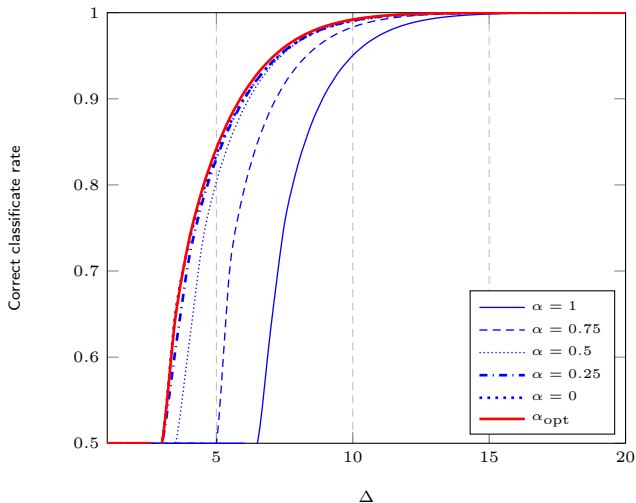
- ▶ eigenvectors are “noisy staircase vectors”
- ▶ conjectured Gaussian fluctuations of eigenvector entries



**Analysis of eigenvectors** reveals:

- ▶ eigenvectors are “noisy staircase vectors”
- ▶ conjectured Gaussian fluctuations of eigenvector entries
- ▶ for  $q_i = q_0$  (homogeneous case), same variance for all entries in same class
- ▶ in non-homogeneous case, we can compute “average variance per class”  
⇒ Heuristic asymptotic performance upper-bound using EM.

## Theoretical Performance Results (uniform distribution for $q_i$ )



**Figure:** Theoretical probability of correct recovery for  $n = 2000$ ,  $K = 2$ ,  $c_1 = 0.6$ ,  $c_2 = 0.4$ ,  $\mu$  uniformly distributed in  $[0.2, 0.8]$ ,  $M = \Delta I_2$ , for  $\Delta \in [0, 20]$ .

## Theoretical Performance Results (uniform distribution for $q_i$ )

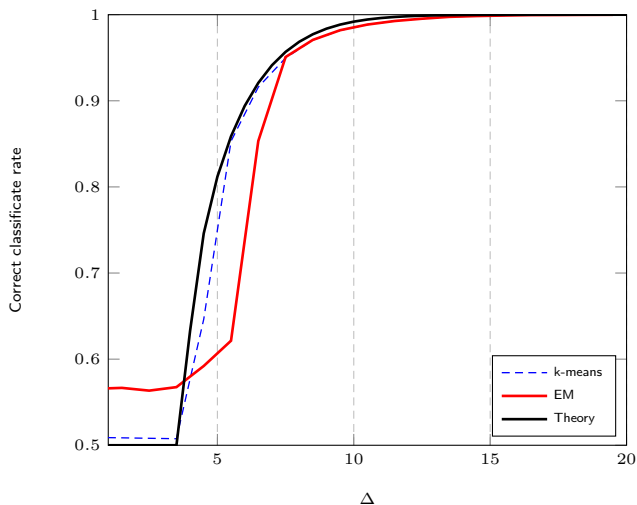


Figure: Probability of correct recovery for  $n = 2000$ ,  $K = 2$ ,  $c_1 = 0.6$ ,  $c_2 = 0.4$ ,  $\mu$  uniformly distributed in  $[0.2, 0.8]$ ,  $M = \Delta I_2$ , for  $\Delta \in [0, 20]$ .

## Some Takeaway messages

**Main findings:**

## Some Takeaway messages

### Main findings:

- ▶ Degree heterogeneity breaks community structures in eigenvectors.  
⇒ Compensation by  $D^{-1}$  normalization of eigenvectors.

# Some Takeaway messages

## Main findings:

- ▶ Degree heterogeneity breaks community structures in eigenvectors.  
⇒ Compensation by  $D^{-1}$  normalization of eigenvectors.
- ▶ Classical debate over “best normalization” of adjacency (or modularity) matrix  $A$  not trivial to solve.  
⇒ With heterogeneous degrees, we found a good on-line method.

# Some Takeaway messages

## Main findings:

- ▶ Degree heterogeneity breaks community structures in eigenvectors.  
⇒ Compensation by  $D^{-1}$  normalization of eigenvectors.
- ▶ Classical debate over “best normalization” of adjacency (or modularity) matrix  $A$  not trivial to solve.  
⇒ With heterogeneous degrees, we found a good on-line method.
- ▶ Simulations support good performances even for “rather sparse” settings.

# Some Takeaway messages

## Main findings:

- ▶ Degree heterogeneity breaks community structures in eigenvectors.  
⇒ Compensation by  $D^{-1}$  normalization of eigenvectors.
- ▶ Classical debate over “best normalization” of adjacency (or modularity) matrix  $A$  not trivial to solve.  
⇒ With heterogeneous degrees, we found a good on-line method.
- ▶ Simulations support good performances even for “rather sparse” settings.

## But strong limitations:



# Some Takeaway messages

## Main findings:

- ▶ Degree heterogeneity breaks community structures in eigenvectors.  
⇒ Compensation by  $D^{-1}$  normalization of eigenvectors.
- ▶ Classical debate over “best normalization” of adjacency (or modularity) matrix  $A$  not trivial to solve.  
⇒ With heterogeneous degrees, we found a good on-line method.
- ▶ Simulations support good performances even for “rather sparse” settings.

## But strong limitations:

- ▶ Key assumption:  $C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$ .  
⇒ Everything collapses if different regime.

# Some Takeaway messages

## Main findings:

- ▶ Degree heterogeneity breaks community structures in eigenvectors.  
⇒ Compensation by  $D^{-1}$  normalization of eigenvectors.
- ▶ Classical debate over “best normalization” of adjacency (or modularity) matrix  $A$  not trivial to solve.  
⇒ With heterogeneous degrees, we found a good on-line method.
- ▶ Simulations support good performances even for “rather sparse” settings.

## But strong limitations:

- ▶ Key assumption:  $C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$ .  
⇒ Everything collapses if different regime.
- ▶ Simulations on small networks in fact give ridiculous arbitrary results.

# Some Takeaway messages

## Main findings:

- ▶ Degree heterogeneity breaks community structures in eigenvectors.  
⇒ Compensation by  $D^{-1}$  normalization of eigenvectors.
- ▶ Classical debate over “best normalization” of adjacency (or modularity) matrix  $A$  not trivial to solve.  
⇒ With heterogeneous degrees, we found a good on-line method.
- ▶ Simulations support good performances even for “rather sparse” settings.

## But strong limitations:

- ▶ Key assumption:  $C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}}$ .  
⇒ Everything collapses if different regime.
- ▶ Simulations on small networks in fact give ridiculous arbitrary results.
- ▶ When is sparse sparse and dense dense?
  - ▶ in theory,  $d_i = O(\log(n))$  is dense...
  - ▶ in practice, assuming dense regime, eigenvalues smear beyond support edges in critical scenarios.

Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

**Kernel Spectral Clustering**

Kernel Spectral Clustering: Subspace Clustering

Semi-supervised Learning

Support Vector Machines

Neural Networks: Extreme Learning Machines

Random Matrices and Robust Estimation

Perspectives

## Problem Statement

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

# Kernel Spectral Clustering

## Problem Statement

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

$$(\text{RatioCut}) \quad \operatorname{argmin}_{\mathcal{S}_1 \cup \dots \cup \mathcal{S}_k = \{1, \dots, n\}} \sum_{i=1}^k \sum_{\substack{j \in \mathcal{S}_i \\ j \notin \mathcal{S}_i}} \frac{\kappa(x_j, x_{\bar{j}})}{|\mathcal{S}_i|}$$

for some similarity kernel  $\kappa(x, y) \geq 0$  (large if  $x$  similar to  $y$ ).

# Kernel Spectral Clustering

## Problem Statement

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

$$(\text{RatioCut}) \operatorname{argmin}_{\mathcal{S}_1 \cup \dots \cup \mathcal{S}_k = \{1, \dots, n\}} \sum_{i=1}^k \sum_{\substack{j \in \mathcal{S}_i \\ j \notin \mathcal{S}_i}} \frac{\kappa(x_j, x_{\bar{j}})}{|\mathcal{S}_i|}$$

for some similarity kernel  $\kappa(x, y) \geq 0$  (large if  $x$  similar to  $y$ ).

- ▶ Can be shown equivalent to

$$(\text{RatioCut}) \operatorname{argmin}_{M \in \mathcal{M}} \operatorname{tr} M^T (D - K) M$$

where  $\mathcal{M} \subset \mathbb{R}^{n \times k} \cap \left\{ M; M_{ij} \in \{0, |\mathcal{S}_j|^{-\frac{1}{2}}\} \right\}$  (in particular,  $M^T M = I_k$ ) and

$$K = \{\kappa(x_i, x_j)\}_{i,j=1}^n, \quad D_{ii} = \sum_{j=1}^n K_{ij}.$$

# Kernel Spectral Clustering

## Problem Statement

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

$$(\text{RatioCut}) \operatorname{argmin}_{\mathcal{S}_1 \cup \dots \cup \mathcal{S}_k = \{1, \dots, n\}} \sum_{i=1}^k \sum_{\substack{j \in \mathcal{S}_i \\ j \notin \mathcal{S}_i}} \frac{\kappa(x_j, x_{\bar{j}})}{|\mathcal{S}_i|}$$

for some similarity kernel  $\kappa(x, y) \geq 0$  (large if  $x$  similar to  $y$ ).

- ▶ Can be shown equivalent to

$$(\text{RatioCut}) \operatorname{argmin}_{M \in \mathcal{M}} \operatorname{tr} M^T (D - K) M$$

where  $\mathcal{M} \subset \mathbb{R}^{n \times k} \cap \left\{ M; M_{ij} \in \{0, |\mathcal{S}_j|^{-\frac{1}{2}}\} \right\}$  (in particular,  $M^T M = I_k$ ) and

$$K = \{\kappa(x_i, x_j)\}_{i,j=1}^n, \quad D_{ii} = \sum_{j=1}^n K_{ij}.$$

- ▶ But **integer problem!** Usually NP-complete.



## Towards kernel spectral clustering

- ▶ Kernel spectral clustering: **discrete-to-continuous relaxations** of such metrics

$$(\text{RatioCut}) \quad \operatorname{argmin}_{M, M^T M = I_K} \operatorname{tr} M^T (D - K) M$$

i.e., eigenvector problem:

1. find eigenvectors of smallest eigenvalues
2. retrieve classes from eigenvector components

## Towards kernel spectral clustering

- ▶ Kernel spectral clustering: **discrete-to-continuous relaxations** of such metrics

$$(\text{RatioCut}) \quad \operatorname{argmin}_{M, M^T M = I_K} \operatorname{tr} M^T (D - K) M$$

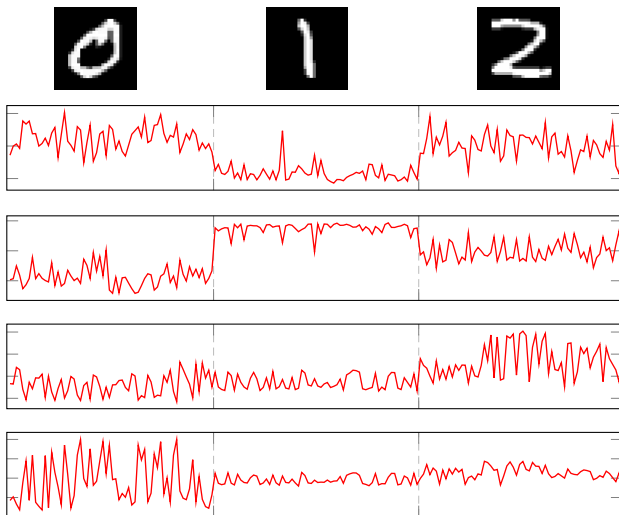
i.e., eigenvector problem:

1. find eigenvectors of smallest eigenvalues
2. retrieve classes from eigenvector components

- ▶ Refinements:

- ▶ working on  $K$ ,  $D - K$ ,  $I_n - D^{-1}K$ ,  $I_n - D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ , etc.
- ▶ several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

# Kernel Spectral Clustering



**Figure:** Leading four eigenvectors of  $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$  for MNIST data.

## Current state:

- ▶ Algorithms derived from ad-hoc procedures (e.g., relaxation).
- ▶ Little understanding of performance, even for Gaussian mixtures!
- ▶ Let alone when both  $p$  and  $n$  are large (BigData setting)

# Methodology and objectives

## Current state:

- ▶ Algorithms derived from ad-hoc procedures (e.g., relaxation).
- ▶ Little understanding of performance, even for Gaussian mixtures!
- ▶ Let alone when **both  $p$  and  $n$  are large (BigData setting)**

## Objectives and Roadmap:

- ▶ Develop **mathematical analysis framework** for BigData kernel spectral clustering  
( $p, n \rightarrow \infty$ )

# Methodology and objectives

## Current state:

- ▶ Algorithms derived from ad-hoc procedures (e.g., relaxation).
- ▶ Little understanding of performance, even for Gaussian mixtures!
- ▶ Let alone when **both  $p$  and  $n$  are large (BigData setting)**

## Objectives and Roadmap:

- ▶ Develop **mathematical analysis framework** for BigData kernel spectral clustering ( $p, n \rightarrow \infty$ )
- ▶ Understand:
  1. Phase transition effects (i.e., when is clustering possible?)
  2. Content of each eigenvector
  3. Influence of kernel function
  4. Performance comparison of clustering algorithms

# Methodology and objectives

## Current state:

- ▶ Algorithms derived from ad-hoc procedures (e.g., relaxation).
- ▶ Little understanding of performance, even for Gaussian mixtures!
- ▶ Let alone when **both  $p$  and  $n$  are large (BigData setting)**

## Objectives and Roadmap:

- ▶ Develop **mathematical analysis framework** for BigData kernel spectral clustering ( $p, n \rightarrow \infty$ )
- ▶ Understand:
  1. Phase transition effects (i.e., when is clustering possible?)
  2. Content of each eigenvector
  3. Influence of kernel function
  4. Performance comparison of clustering algorithms

## Methodology:

- ▶ Use statistical assumptions (Gaussian mixture)
- ▶ Benefit from doubly-infinite independence and **random matrix tools**

# Model and Assumptions

## Gaussian mixture model:

- ▶  $x_1, \dots, x_n \in \mathbb{R}^p$ ,
- ▶  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ ,
- ▶  $x_1, \dots, x_{n_1} \in \mathcal{C}_1, \dots, x_{n-n_k+1}, \dots, x_n \in \mathcal{C}_k$ ,
- ▶  $\mathcal{C}_a = \{x \mid x \sim \mathcal{N}(\mu_a, C_a)\}$ .

Then, for  $x_i \in \mathcal{C}_a$ , with  $w_i \sim N(0, C_a)$ ,

$$x_i = \mu_a + w_i.$$



# Model and Assumptions

## Gaussian mixture model:

- ▶  $x_1, \dots, x_n \in \mathbb{R}^p$ ,
- ▶  $k$  classes  $\mathcal{C}_1, \dots, \mathcal{C}_k$ ,
- ▶  $x_1, \dots, x_{n_1} \in \mathcal{C}_1, \dots, x_{n-n_k+1}, \dots, x_n \in \mathcal{C}_k$ ,
- ▶  $\mathcal{C}_a = \{x \mid x \sim \mathcal{N}(\mu_a, C_a)\}$ .

Then, for  $x_i \in \mathcal{C}_a$ , with  $w_i \sim N(0, C_a)$ ,

$$x_i = \mu_a + w_i.$$

[Convergence Rate] As  $n \rightarrow \infty$ ,

1. **Data scaling:**  $\frac{p}{n} \rightarrow c_0 \in (0, \infty)$ ,
2. **Class scaling:**  $\frac{n_a}{n} \rightarrow c_a \in (0, 1)$ ,
3. **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

$$\|\mu_a^\circ\| = O(1)$$

4. **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

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

## 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  $f$ .

## 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  $f$ .

- ▶ We study the normalized recentered Laplacian:

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

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

**Difficulty:**  $L$  is a very intractable random matrix

- ▶ non-linear  $f$
- ▶ non-trivial dependence between entries of  $L$

**Difficulty:**  $L$  is a very intractable random matrix

- ▶ non-linear  $f$
- ▶ non-trivial dependence between entries of  $L$

**Strategy:**

1. Find random equivalent  $\hat{L}$  (i.e.,  $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$  as  $n, p \rightarrow \infty$ ) based on:
  - ▶ **concentration:**  $K_{ij} \rightarrow \text{constant}$  as  $n, p \rightarrow \infty$  (for all  $i \neq j$ )
  - ▶ Taylor expansion around limit point

**Difficulty:**  $L$  is a very intractable random matrix

- ▶ non-linear  $f$
- ▶ non-trivial dependence between entries of  $L$

**Strategy:**

1. Find random equivalent  $\hat{L}$  (i.e.,  $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$  as  $n, p \rightarrow \infty$ ) based on:
  - ▶ **concentration:**  $K_{ij} \rightarrow \text{constant}$  as  $n, p \rightarrow \infty$  (for all  $i \neq j$ )
  - ▶ Taylor expansion around limit point
2. Apply **spiked random matrix approach** to study:
  - ▶ existence of isolated eigenvalues in  $\hat{L}$ : **phase transition**

**Difficulty:**  $L$  is a very intractable random matrix

- ▶ non-linear  $f$
- ▶ non-trivial dependence between entries of  $L$

**Strategy:**

1. Find random equivalent  $\hat{L}$  (i.e.,  $\|L - \hat{L}\| \xrightarrow{\text{a.s.}} 0$  as  $n, p \rightarrow \infty$ ) based on:
  - ▶ **concentration:**  $K_{ij} \rightarrow \text{constant}$  as  $n, p \rightarrow \infty$  (for all  $i \neq j$ )
  - ▶ Taylor expansion around limit point
2. Apply **spiked random matrix approach** to study:
  - ▶ existence of isolated eigenvalues in  $\hat{L}$ : **phase transition**
  - ▶ eigenvector projections on canonical class-basis

# Random Matrix Equivalent

Results on  $K$ :

- **Key Remark:** Under our assumptions, uniformly on  $i, j \in \{1, \dots, n\}$ ,

$$\frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

for some **common limit**  $\tau$ .



# Random Matrix Equivalent

## Results on $K$ :

- **Key Remark:** Under our assumptions, uniformly on  $i, j \in \{1, \dots, n\}$ ,

$$\frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

for some **common limit**  $\tau$ .

- large dimensional approximation for  $K$ :

$$K = \underbrace{f(\tau)1_n 1_n^T}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}A_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{A_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

# Random Matrix Equivalent

## Results on $K$ :

- **Key Remark:** Under our assumptions, uniformly on  $i, j \in \{1, \dots, n\}$ ,

$$\frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

for some **common limit**  $\tau$ .

- large dimensional approximation for  $K$ :

$$K = \underbrace{f(\tau)1_n 1_n^T}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}A_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{A_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

- **difficult to handle** (3 orders to manipulate!)

# Random Matrix Equivalent

## Results on $K$ :

- ▶ **Key Remark:** Under our assumptions, uniformly on  $i, j \in \{1, \dots, n\}$ ,

$$\frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

for some **common limit**  $\tau$ .

- ▶ large dimensional approximation for  $K$ :

$$K = \underbrace{f(\tau)1_n 1_n^T}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}A_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{A_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

- ▶ **difficult to handle** (3 orders to manipulate!)

## Observation: Spectrum of $L$ :

- ▶ Dominant eigenvalue  $n$  with eigenvector  $D^{\frac{1}{2}} 1_n$
- ▶ **All other eigenvalues of order  $O(1)$ .**

# Random Matrix Equivalent

## Results on $K$ :

- ▶ **Key Remark:** Under our assumptions, uniformly on  $i, j \in \{1, \dots, n\}$ ,

$$\frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$$

for some **common limit**  $\tau$ .

- ▶ large dimensional approximation for  $K$ :

$$K = \underbrace{f(\tau)1_n 1_n^\top}_{O_{\|\cdot\|}(n)} + \underbrace{\sqrt{n}A_1}_{\text{low rank, } O_{\|\cdot\|}(\sqrt{n})} + \underbrace{A_2}_{\text{informative terms, } O_{\|\cdot\|}(1)}$$

- ▶ **difficult to handle** (3 orders to manipulate!)

## Observation: Spectrum of $L$ :

- ▶ Dominant eigenvalue  $n$  with eigenvector  $D^{\frac{1}{2}}1_n$
- ▶ **All other eigenvalues of order  $O(1)$ .**

⇒ Naturally leads to study:

- ▶ Projected normalized Laplacian (or “modularity”-type Laplacian):

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

- ▶ Dominant (normalized) eigenvector  $\frac{D^{\frac{1}{2}}1_n}{\sqrt{1_n^\top D 1_n}}$ .

# Random Matrix Equivalent

## Theorem (Random Matrix Equivalent)

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

$$\hat{L}' = -2 \frac{f'(\tau)}{f(\tau)} \left[ \frac{1}{p} P W^\top W P + U B U^\top \right] + \alpha(\tau) I_n$$

and  $\tau = \frac{2}{p} \text{tr} C^\circ$ ,  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$ ,

$$U = \left[ \frac{1}{\sqrt{p}} J, \Phi, \psi \right] \in \mathbb{R}^{n \times (2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - \mathbf{1}_k c^\top & \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t \\ I_k - c \mathbf{1}_k^\top & 0_{k \times k} & 0_{k \times 1} \\ \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t^\top & 0_{1 \times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4) \times (2k+4)}$$

$$B_{11} = M^\top M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t t^\top - \frac{f''(\tau)}{f'(\tau)} T + \frac{p}{n} \frac{f(\tau) \alpha(\tau)}{2f'(\tau)} \mathbf{1}_k \mathbf{1}_k^\top \in \mathbb{R}^{k \times k}.$$

# Random Matrix Equivalent

## Theorem (Random Matrix Equivalent)

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

$$\hat{L}' = -2 \frac{f'(\tau)}{f(\tau)} \left[ \frac{1}{p} P W^\top W P + U B U^\top \right] + \alpha(\tau) I_n$$

and  $\tau = \frac{2}{p} \text{tr} C^\circ$ ,  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} 1_n 1_n^\top$ ,

$$U = \left[ \frac{1}{\sqrt{p}} J, \Phi, \psi \right] \in \mathbb{R}^{n \times (2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - 1_k c^\top & \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t \\ I_k - c 1_k^\top & 0_{k \times k} & 0_{k \times 1} \\ \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t^\top & 0_{1 \times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4) \times (2k+4)}$$

$$B_{11} = M^\top M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t t^\top - \frac{f''(\tau)}{f'(\tau)} T + \frac{p}{n} \frac{f(\tau) \alpha(\tau)}{2f'(\tau)} 1_k 1_k^\top \in \mathbb{R}^{k \times k}.$$

**Important Notations:**

$\frac{1}{\sqrt{p}} J = [j_1, \dots, j_k] \in \mathbb{R}^{n \times k}$ ,  $j_a$  canonical vector of class  $\mathcal{C}_a$ .

# Random Matrix Equivalent

## Theorem (Random Matrix Equivalent)

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

$$\hat{L}' = -2 \frac{f'(\tau)}{f(\tau)} \left[ \frac{1}{p} P W^\top W P + U B U^\top \right] + \alpha(\tau) I_n$$

and  $\tau = \frac{2}{p} \text{tr} C^\circ$ ,  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$ ,

$$U = \left[ \frac{1}{\sqrt{p}} J, \Phi, \psi \right] \in \mathbb{R}^{n \times (2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - \mathbf{1}_k c^\top & \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t \\ I_k - c \mathbf{1}_k^\top & 0_{k \times k} & 0_{k \times 1} \\ \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t^\top & 0_{1 \times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4) \times (2k+4)}$$

$$B_{11} = M^\top M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t t^\top - \frac{f''(\tau)}{f'(\tau)} T + \frac{p}{n} \frac{f(\tau) \alpha(\tau)}{2f'(\tau)} \mathbf{1}_k \mathbf{1}_k^\top \in \mathbb{R}^{k \times k}.$$

### Important Notations:

$$M = [\mu_1^\circ, \dots, \mu_k^\circ] \in \mathbb{R}^{n \times k}, \mu_a^\circ = \mu_a - \sum_{b=1}^k \frac{n_b}{n} \mu_b.$$

# Random Matrix Equivalent

## Theorem (Random Matrix Equivalent)

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

$$\hat{L}' = -2 \frac{f'(\tau)}{f(\tau)} \left[ \frac{1}{p} P W^T W P + U B U^T \right] + \alpha(\tau) I_n$$

and  $\tau = \frac{2}{p} \text{tr} C^\circ$ ,  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ ,

$$U = \left[ \frac{1}{\sqrt{p}} J, \Phi, \psi \right] \in \mathbb{R}^{n \times (2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - \mathbf{1}_k c^T & \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t \\ I_k - c \mathbf{1}_k^T & 0_{k \times k} & 0_{k \times 1} \\ \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t^T & 0_{1 \times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4) \times (2k+4)}$$

$$B_{11} = M^T M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t t^T - \frac{f''(\tau)}{f'(\tau)} T + \frac{p}{n} \frac{f(\tau) \alpha(\tau)}{2f'(\tau)} \mathbf{1}_k \mathbf{1}_k^T \in \mathbb{R}^{k \times k}.$$

**Important Notations:**

$$t = \left[ \frac{1}{\sqrt{p}} \text{tr} C_1^\circ, \dots, \frac{1}{\sqrt{p}} \text{tr} C_k^\circ \right] \in \mathbb{R}^k, \quad C_a^\circ = C_a - \sum_{b=1}^k \frac{n_b}{n} C_b.$$



# Random Matrix Equivalent

## Theorem (Random Matrix Equivalent)

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

$$\hat{L}' = -2 \frac{f'(\tau)}{f(\tau)} \left[ \frac{1}{p} P W^T W P + U B U^T \right] + \alpha(\tau) I_n$$

and  $\tau = \frac{2}{p} \text{tr} C^\circ$ ,  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} 1_n 1_n^T$ ,

$$U = \left[ \frac{1}{\sqrt{p}} J, \Phi, \psi \right] \in \mathbb{R}^{n \times (2k+4)}$$

$$B = \begin{bmatrix} B_{11} & I_k - 1_k c^T & \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t \\ I_k - c 1_k^T & 0_{k \times k} & 0_{k \times 1} \\ \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t^T & 0_{1 \times k} & \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \end{bmatrix} \in \mathbb{R}^{(2k+4) \times (2k+4)}$$

$$B_{11} = M^T M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t t^T - \frac{f''(\tau)}{f'(\tau)} T + \frac{p}{n} \frac{f(\tau) \alpha(\tau)}{2f'(\tau)} 1_k 1_k^T \in \mathbb{R}^{k \times k}.$$

**Important Notations:**

$$T = \left\{ \frac{1}{p} \text{tr} C_a^\circ C_b^\circ \right\}_{a,b=1}^k \in \mathbb{R}^{k \times k}, \quad C_a^\circ = C_a - \sum_{b=1}^k \frac{n_b}{n} C_b.$$

## Some consequences:

- ▶  $\hat{L}'$  is a spiked model:  $UBU^T$  seen as low rank perturbation of  $\frac{1}{p}PW^TW P$

## Some consequences:

- ▶  $\hat{L}'$  is a spiked model:  $UBU^\top$  seen as low rank perturbation of  $\frac{1}{p}PW^\top WP$
- ▶ If  $f'(\tau) = 0$ ,
  - ▶  $L'$  asymptotically deterministic!
  - ▶ only  $t$  and  $T$  can be discriminated upon
- ▶ If  $f''(\tau) = 0$ , (e.g.,  $f(x) = x$ )  $T$  unused
- ▶ If  $\frac{5f'(\tau)}{8f(\tau)} = \frac{f''(\tau)}{2f'(\tau)}$ ,  $t$  (seemingly) unused

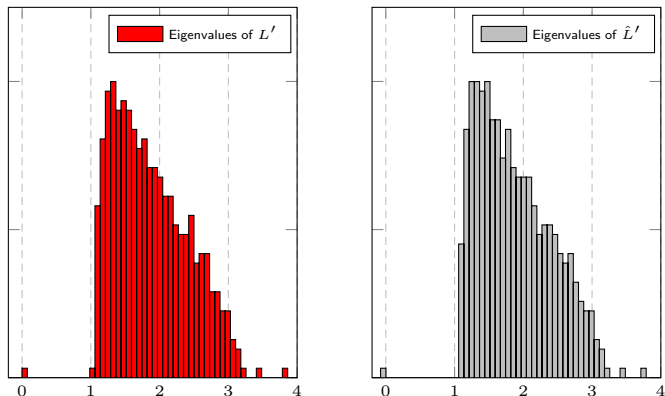
## Some consequences:

- ▶  $\hat{L}'$  is a spiked model:  $UBU^T$  seen as low rank perturbation of  $\frac{1}{p}PW^TWP$
- ▶ If  $f'(\tau) = 0$ ,
  - ▶  $L'$  asymptotically deterministic!
  - ▶ only  $t$  and  $T$  can be discriminated upon
- ▶ If  $f''(\tau) = 0$ , (e.g.,  $f(x) = x$ )  $T$  unused
- ▶ If  $\frac{5f'(\tau)}{8f(\tau)} = \frac{f''(\tau)}{2f'(\tau)}$ ,  $t$  (seemingly) unused

## Further analysis:

- ▶ Determine separability condition for eigenvalues
- ▶ Evaluate eigenvalue positions when separable
- ▶ Evaluate eigenvector projection to canonical basis  $j_1, \dots, j_k$
- ▶ Evaluate fluctuation of eigenvectors.

## Isolated eigenvalues: Gaussian inputs



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

# Theoretical Findings versus MNIST

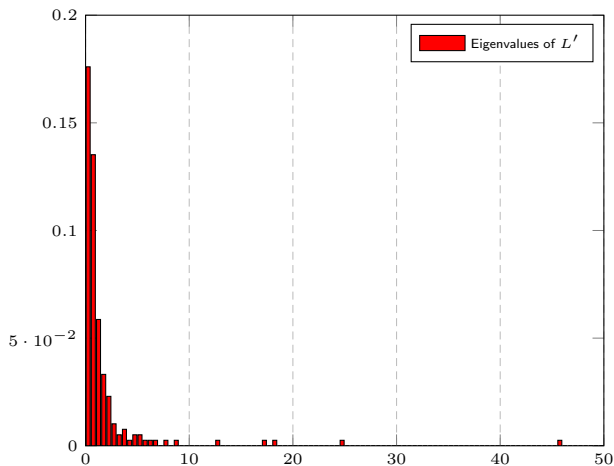


Figure: Eigenvalues of  $L'$  (red) and (equivalent Gaussian model)  $\hat{L}'$  (white), MNIST data,  $p = 784$ ,  $n = 192$ .

# Theoretical Findings versus MNIST

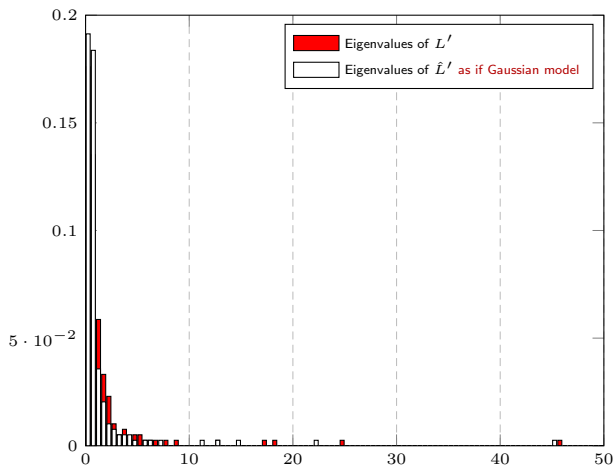


Figure: Eigenvalues of  $L'$  (red) and (equivalent Gaussian model)  $\hat{L}'$  (white), MNIST data,  $p = 784$ ,  $n = 192$ .

# Theoretical Findings versus MNIST

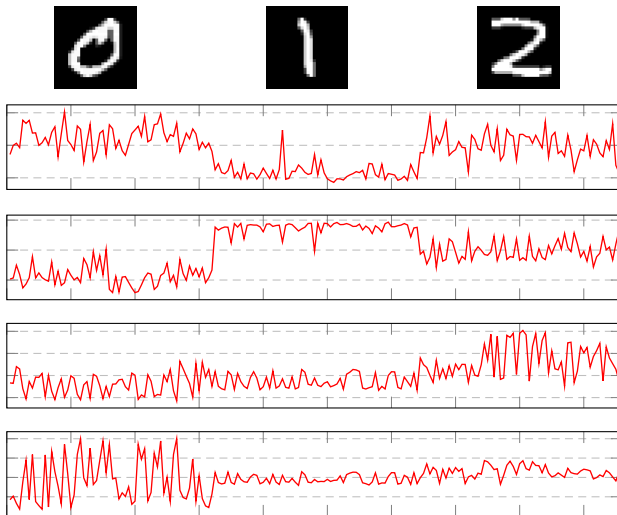


Figure: Leading four eigenvectors of  $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$  for MNIST data (red) and theoretical findings (blue).



# Theoretical Findings versus MNIST

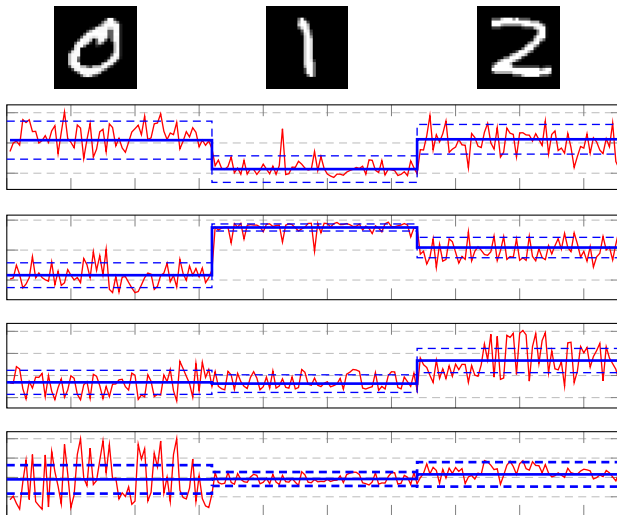
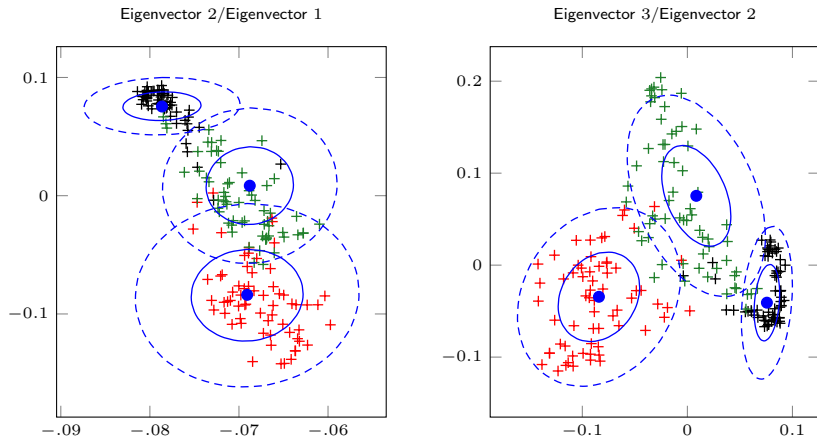


Figure: Leading four eigenvectors of  $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$  for MNIST data (red) and theoretical findings (blue).

# Theoretical Findings versus MNIST



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

## Further Results and Some Takeaway messages

**General surprising findings:**

## Further Results and Some Takeaway messages

### General surprising findings:

- ▶ “Good kernel functions”  $f$  need not be decreasing.

## Further Results and Some Takeaway messages

### General surprising findings:

- ▶ “Good kernel functions”  $f$  need not be decreasing.
- ▶ Dominant parameters in large dimensions are first three derivatives at  $\tau$ .

# Further Results and Some Takeaway messages

## General surprising findings:

- ▶ “Good kernel functions”  $f$  need not be decreasing.
- ▶ Dominant parameters in large dimensions are first three derivatives at  $\tau$ .
- ▶ Clustering possible despite  $\|x_i - x_j\|^2 \rightarrow \tau$ , i.e., no first order data difference  
⇒ **Breaks original intuitions and problem layout!**

# Further Results and Some Takeaway messages

## General surprising findings:

- ▶ “Good kernel functions”  $f$  need not be decreasing.
- ▶ Dominant parameters in large dimensions are first three derivatives at  $\tau$ .
- ▶ Clustering possible despite  $\|x_i - x_j\|^2 \rightarrow \tau$ , i.e., no first order data difference  
⇒ **Breaks original intuitions and problem layout!**

## Further surprises. . . :

# Further Results and Some Takeaway messages

## General surprising findings:

- ▶ “Good kernel functions”  $f$  need not be decreasing.
- ▶ Dominant parameters in large dimensions are first three derivatives at  $\tau$ .
- ▶ Clustering possible despite  $\|x_i - x_j\|^2 \rightarrow \tau$ , i.e., no first order data difference  
⇒ **Breaks original intuitions and problem layout!**

## Further surprises. . . :

- ▶ For  $C_1 = \dots = C_K = I_p$ , **kernel choice is irrelevant!** (as long as  $f'(\tau) \neq 0$ )



# Further Results and Some Takeaway messages

## General surprising findings:

- ▶ “Good kernel functions”  $f$  need not be decreasing.
- ▶ Dominant parameters in large dimensions are first three derivatives at  $\tau$ .
- ▶ Clustering possible despite  $\|x_i - x_j\|^2 \rightarrow \tau$ , i.e., no first order data difference  
 $\Rightarrow$  **Breaks original intuitions and problem layout!**

## Further surprises...:

- ▶ For  $C_1 = \dots = C_K = I_p$ , **kernel choice is irrelevant!** (as long as  $f'(\tau) \neq 0$ )
- ▶ For  $\mu_1 = \dots = \mu_K = 0$  and  $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$ , **only ONE isolated eigenvector!**

# Further Results and Some Takeaway messages

## General surprising findings:

- ▶ “Good kernel functions”  $f$  need not be decreasing.
- ▶ Dominant parameters in large dimensions are first three derivatives at  $\tau$ .
- ▶ Clustering possible despite  $\|x_i - x_j\|^2 \rightarrow \tau$ , i.e., no first order data difference  
 $\Rightarrow$  **Breaks original intuitions and problem layout!**

## Further surprises...:

- ▶ For  $C_1 = \dots = C_K = I_p$ , **kernel choice is irrelevant!** (as long as  $f'(\tau) \neq 0$ )
- ▶ For  $\mu_1 = \dots = \mu_K = 0$  and  $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$ , **only ONE isolated eigenvector!**
- ▶ It is possible to observe **irrelevant eigenvectors!** (that contain only noise)

# Further Results and Some Takeaway messages

## General surprising findings:

- ▶ “Good kernel functions”  $f$  **need not be decreasing**.
- ▶ Dominant parameters in large dimensions are **first three derivatives at  $\tau$** .
- ▶ Clustering possible despite  $\|x_i - x_j\|^2 \rightarrow \tau$ , i.e., **no first order data difference**  
 $\Rightarrow$  **Breaks original intuitions and problem layout!**

## Further surprises. . . :

- ▶ For  $C_1 = \dots = C_K = I_p$ , **kernel choice is irrelevant!** (as long as  $f'(\tau) \neq 0$ )
- ▶ For  $\mu_1 = \dots = \mu_K = 0$  and  $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$ , **only ONE isolated eigenvector!**
- ▶ It is possible to observe **irrelevant eigenvectors!** (that contain only noise)

## Validity of the Results:

# Further Results and Some Takeaway messages

## General surprising findings:

- ▶ “Good kernel functions”  $f$  **need not be decreasing**.
- ▶ Dominant parameters in large dimensions are **first three derivatives at  $\tau$** .
- ▶ Clustering possible despite  $\|x_i - x_j\|^2 \rightarrow \tau$ , i.e., **no first order data difference**  
 $\Rightarrow$  **Breaks original intuitions and problem layout!**

## Further surprises...:

- ▶ For  $C_1 = \dots = C_K = I_p$ , **kernel choice is irrelevant!** (as long as  $f'(\tau) \neq 0$ )
- ▶ For  $\mu_1 = \dots = \mu_K = 0$  and  $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$ , **only ONE isolated eigenvector!**
- ▶ It is possible to observe **irrelevant eigenvectors!** (that contain only noise)

## Validity of the Results:

- ▶ Needs a concentration of measure assumption:  $\|x_i - x_j\|^2 \rightarrow \tau$ .
- ▶ Invalid for heavy-tailed distributions (where  $\|x_i\| = \|\sqrt{\tau_i} z_i\|$  needs not converge).

# Further Results and Some Takeaway messages

## General surprising findings:

- ▶ “Good kernel functions”  $f$  **need not be decreasing**.
- ▶ Dominant parameters in large dimensions are **first three derivatives at  $\tau$** .
- ▶ Clustering possible despite  $\|x_i - x_j\|^2 \rightarrow \tau$ , i.e., **no first order data difference**  
 $\Rightarrow$  **Breaks original intuitions and problem layout!**

## Further surprises. . . :

- ▶ For  $C_1 = \dots = C_K = I_p$ , **kernel choice is irrelevant!** (as long as  $f'(\tau) \neq 0$ )
- ▶ For  $\mu_1 = \dots = \mu_K = 0$  and  $C_a = (1 + \gamma_a p^{-\frac{1}{2}})I_p$ , **only ONE isolated eigenvector!**
- ▶ It is possible to observe **irrelevant eigenvectors!** (that contain only noise)

## Validity of the Results:

- ▶ Needs a concentration of measure assumption:  $\|x_i - x_j\|^2 \rightarrow \tau$ .
- ▶ Invalid for heavy-tailed distributions (where  $\|x_i\| = \|\sqrt{\tau_i} z_i\|$  needs not converge).
- ▶ **Suprising fit between theory and practice:** are images like Gaussian vectors?
  - ▶ kernels extract primarily first order properties (means, covariances)
  - ▶ without image processing (rotations, scale invariance), good enough features.

Last word: the suprising case  $f'(\tau) = 0...$

Reminder:

### Theorem (Random Matrix Equivalent)

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

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

and  $\tau = \frac{2}{p} \text{tr} C^o$ ,  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ ,

$$U = \left[ \frac{1}{\sqrt{p}} J, * \right], \quad B = \begin{bmatrix} B_{11} & * \\ * & * \end{bmatrix}$$

$$B_{11} = M^T M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t t^T - \frac{f''(\tau)}{f'(\tau)} T + \frac{p}{n} \frac{f(\tau)\alpha(\tau)}{2f'(\tau)} \mathbf{1}_k \mathbf{1}_k^T.$$

## Last word: the suprising case $f'(\tau) = 0...$

Reminder:

### Theorem (Random Matrix Equivalent)

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

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

and  $\tau = \frac{2}{p} \text{tr} C^\circ$ ,  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ ,

$$U = \left[ \frac{1}{\sqrt{p}} J, * \right], \quad B = \begin{bmatrix} B_{11} & * \\ * & * \end{bmatrix}$$

$$B_{11} = M^T M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t t^T - \frac{f''(\tau)}{f'(\tau)} T + \frac{p}{n} \frac{f(\tau)\alpha(\tau)}{2f'(\tau)} \mathbf{1}_k \mathbf{1}_k^T.$$

When  $f'(\tau) \rightarrow 0$ ,

- Means  $M$  disappears  $\Rightarrow$  Impossible classification from means.

## Last word: the surprising case $f'(\tau) = 0 \dots$

Reminder:

### Theorem (Random Matrix Equivalent)

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

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

and  $\tau = \frac{2}{p} \text{tr} C^\circ$ ,  $W = [w_1, \dots, w_n] \in \mathbb{R}^{p \times n}$  ( $x_i = \mu_a + w_i$ ),  $P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$ ,

$$U = \left[ \frac{1}{\sqrt{p}} J, * \right], \quad B = \begin{bmatrix} B_{11} & * \\ * & * \end{bmatrix}$$

$$B_{11} = M^T M + \left( \frac{5f'(\tau)}{8f(\tau)} - \frac{f''(\tau)}{2f'(\tau)} \right) t t^T - \frac{f''(\tau)}{f'(\tau)} T + \frac{p}{n} \frac{f(\tau)\alpha(\tau)}{2f'(\tau)} \mathbf{1}_k \mathbf{1}_k^T.$$

When  $f'(\tau) \rightarrow 0$ ,

- ▶ Means  $M$  disappears  $\Rightarrow$  Impossible classification from means.
- ▶ **More importantly:**  $P W W^T P$  disappears  
 $\Rightarrow$  Asymptotic **deterministic** matrix equivalent!  
 $\Rightarrow$  **Perfect asymptotic clustering in theory!**



Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

Kernel Spectral Clustering

**Kernel Spectral Clustering: Subspace Clustering**

Semi-supervised Learning

Support Vector Machines

Neural Networks: Extreme Learning Machines

Random Matrices and Robust Estimation

Perspectives

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

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

**Method:**

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

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

# Position of the Problem

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

**Method:**

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

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

- ▶ Performance of  $L = nD^{-\frac{1}{2}}KD^{-\frac{1}{2}} - n\frac{D^{\frac{1}{2}}1_n1_n^TD^{\frac{1}{2}}}{1_n^TD1_n}$ , with

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

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

## Model and Reminders

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

## Model and Reminders

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

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

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

## Model and Reminders

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

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

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

### Theorem (Corollary of Previous Section)

Let  $f$  smooth with  $f'(2) \neq 0$ . Then, under Assumptions 1–2a,

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

exhibits *phase transition phenomenon*

## Model and Reminders

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

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

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

### Theorem (Corollary of Previous Section)

Let  $f$  smooth with  $f'(2) \neq 0$ . Then, under Assumptions 1–2a,

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

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

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

has sufficiently large eigenvalues.



## The case $f'(2) = 0$

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

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

## The case $f'(2) = 0$

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

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

(in this regime, *previous kernels clearly fail*)

## The case $f'(2) = 0$

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

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

(in this regime, *previous kernels clearly fail*)

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

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

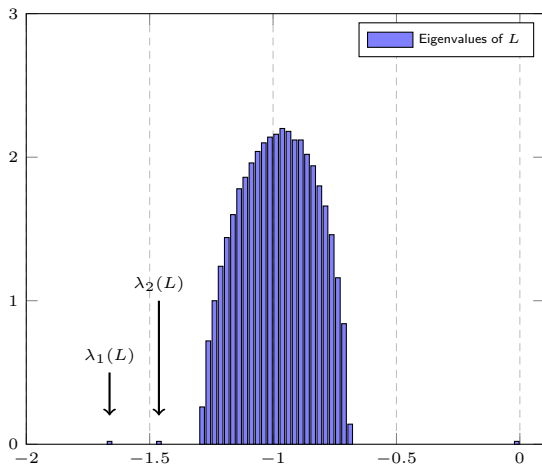
$$\mathcal{L} \equiv \sqrt{p} \frac{f(2)}{2f''(2)} \left[ \textcolor{red}{L} - \frac{f(0) - f(2)}{f(2)} P \right], \quad P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

Then, under Assumptions 1–2b,

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

where  $\Phi_{ij} = \delta_{i \neq j} \sqrt{p} [(x_i^\top x_j)^2 - E[(x_i^\top x_j)^2]]$ .

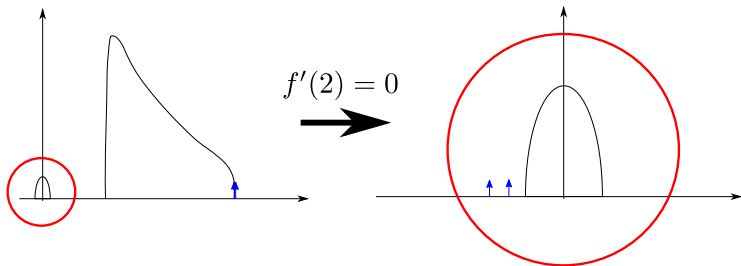
## The case $f'(2) = 0$



**Figure:** Eigenvalues of  $L$ ,  $p = 1000$ ,  $n = 2000$ ,  $k = 3$ ,  $c_1 = c_2 = 1/4$ ,  $c_3 = 1/2$ ,  $C_i \propto I_p + (p/8)^{-\frac{5}{4}} W_i W_i^T$ ,  $W_i \in \mathbb{R}^{p \times (p/8)}$  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  $\Phi$

## The case $f'(2) = 0$

**Roadmap.** We now need to:

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

## The case $f'(2) = 0$

**Roadmap.** We now need to:

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



## The case $f'(2) = 0$

**Roadmap.** We now need to:

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

### Theorem (Semi-circle law for $\Phi$ )

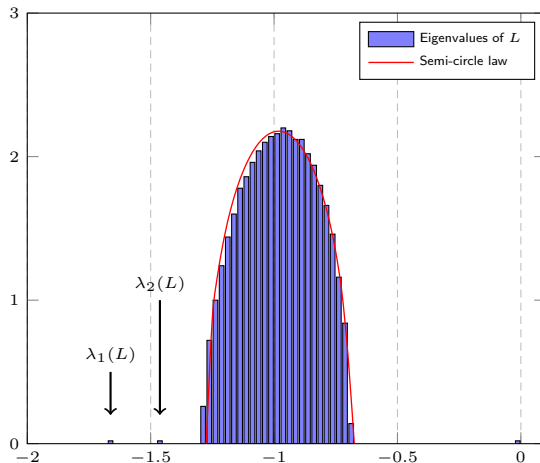
Let  $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(\mathcal{L})}$ . Then, under Assumption 1–2b,

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

with  $\mu$  the semi-circle distribution

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

## The case $f'(2) = 0$



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

## The case $f'(2) = 0$

Denote now

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

## The case $f'(2) = 0$

Denote now

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

### Theorem (Isolated Eigenvalues)

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

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

## The case $f'(2) = 0$

Denote now

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

### Theorem (Isolated Eigenvectors)

For each isolated eigenpair  $(\lambda_i, u_i)$  of  $\mathcal{L}$  corresponding to  $(\nu_i, v_i)$  of  $\mathcal{T}$ , write

$$u_i = \sum_{a=1}^k \alpha_i^a \frac{j_a}{\sqrt{n_a}} + \sigma_i^a w_i^a$$

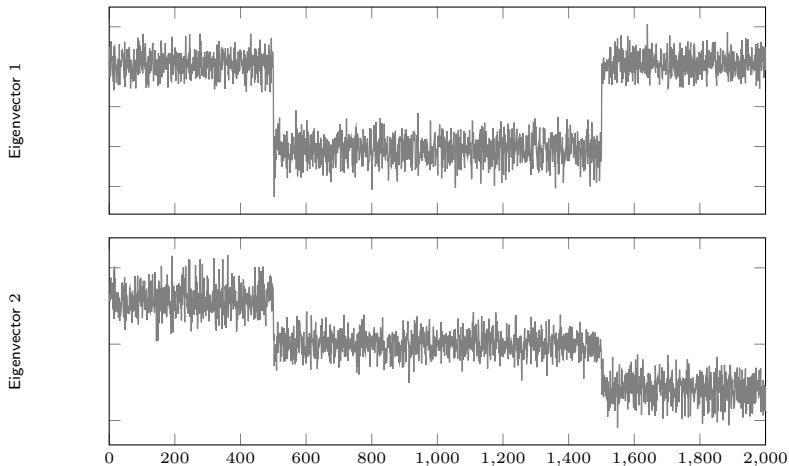
with  $j_a = [0_{n_1}^\top, \dots, 1_{n_a}^\top, \dots, 0_{n_k}^\top]^\top$ ,  $(w_i^a)^\top j_a = 0$ ,  $\text{supp}(w_i^a) = \text{supp}(j_a)$ ,  $\|w_i^a\| = 1$ .  
Then, under Assumptions 1-2b,

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

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

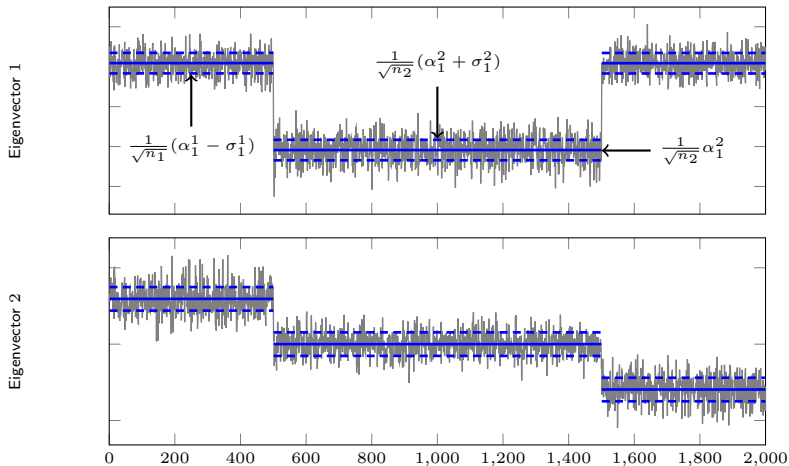
and the fluctuations of  $u_i, u_j$ ,  $i \neq j$ , are asymptotically uncorrelated.

## The case $f'(2) = 0$



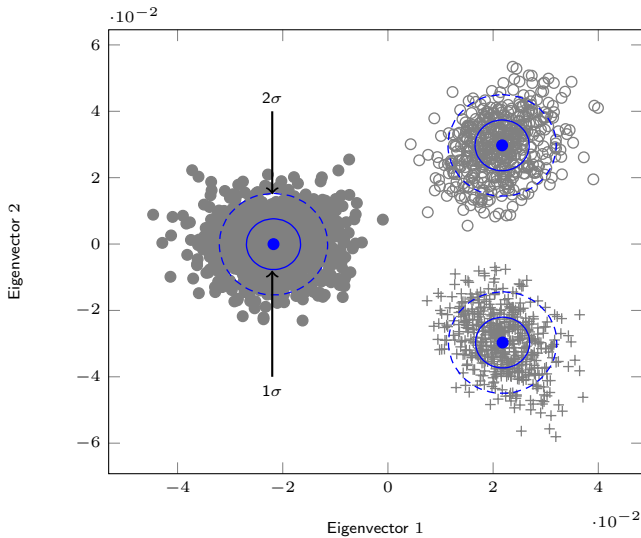
**Figure:** Leading two eigenvectors of  $\mathcal{L}$  (or equivalently of  $L$ ) versus deterministic approximations of  $\alpha_i^a \pm \sigma_i^a$ .

## 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  $\alpha_i^a \pm \sigma_i^a$ .



Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

Kernel Spectral Clustering

Kernel Spectral Clustering: Subspace Clustering

**Semi-supervised Learning**

Support Vector Machines

Neural Networks: Extreme Learning Machines

Random Matrices and Robust Estimation

Perspectives

# Problem Statement

**Context:** Similar to clustering:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k$  classes, but with **labelled** and **unlabelled** data.

# Problem Statement

**Context:** Similar to clustering:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k$  classes, but with **labelled** and **unlabelled** data.
- ▶ Problem statement:  $(d_i = [K1_n]_i)$

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^k \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

such that  $F_{ia} = \delta_{\{x_i \in \mathcal{C}_a\}}$ , for all labelled  $x_i$ .

# Problem Statement

**Context:** Similar to clustering:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k$  classes, but with **labelled** and **unlabelled** data.
- ▶ Problem statement:  $(d_i = [K1_n]_i)$

$$F = \operatorname{argmin}_{F \in \mathbb{R}^{n \times k}} \sum_{a=1}^k \sum_{i,j} K_{ij} (F_{ia} d_i^{\alpha-1} - F_{ja} d_j^{\alpha-1})^2$$

such that  $F_{ia} = \delta_{\{x_i \in \mathcal{C}_a\}}$ , for all labelled  $x_i$ .

- ▶ **Solution:** denoting  $F^{(u)} \in \mathbb{R}^{n_u \times k}$ ,  $F^{(l)} \in \mathbb{R}^{n_l \times k}$  the restriction to unlabelled/labelled data,

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

where we naturally decompose

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

**Using**  $F^{(u)}$ :

- From  $F^{(u)}$ , classification algorithm:

$$\text{Classify } x_i \text{ in } \mathcal{C}_a \Leftrightarrow F_{ia} = \max_{b \in \{1, \dots, k\}} \{F_{ib}\}.$$

Using  $F^{(u)}$ :

- From  $F^{(u)}$ , classification algorithm:

$$\text{Classify } x_i \text{ in } \mathcal{C}_a \Leftrightarrow F_{ia} = \max_{b \in \{1, \dots, k\}} \{F_{ib}\}.$$

**Objectives:** For  $x_i \sim \mathcal{N}(\mu_a, C_a)$ , and as  $n, p \rightarrow \infty$ , ( $n_u, n_l \rightarrow \infty$  or  $n_u \rightarrow \infty$ ,  $n_l = O(1)$ )

# Problem Statement

Using  $F^{(u)}$ :

- From  $F^{(u)}$ , classification algorithm:

$$\text{Classify } x_i \text{ in } \mathcal{C}_a \Leftrightarrow F_{ia} = \max_{b \in \{1, \dots, k\}} \{F_{ib}\}.$$

**Objectives:** For  $x_i \sim \mathcal{N}(\mu_a, C_a)$ , and as  $n, p \rightarrow \infty$ , ( $n_u, n_l \rightarrow \infty$  or  $n_u \rightarrow \infty$ ,  $n_l = O(1)$ )

- Tractable approximation (in norm) for the vectors  $[F_{(u)}]_{\cdot, a}$ ,  $a = 1, \dots, k$
- **Joint asymptotic behavior of  $[F_{(u)}]_{i, \cdot}$ .**  
⇒ From which classification probability is retrieved.

# Problem Statement

Using  $F^{(u)}$ :

- From  $F^{(u)}$ , classification algorithm:

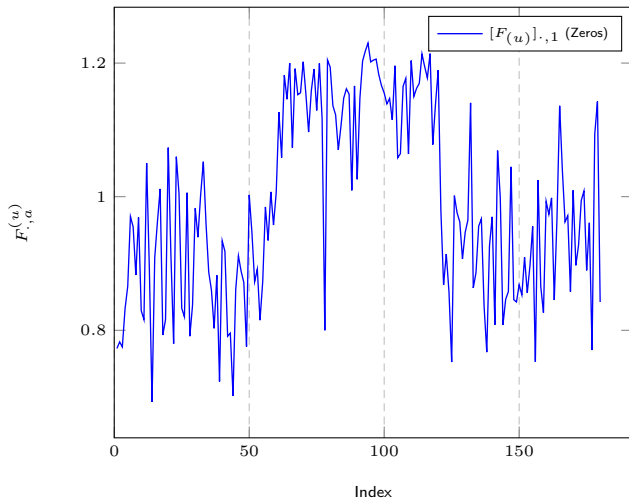
$$\text{Classify } x_i \text{ in } \mathcal{C}_a \Leftrightarrow F_{ia} = \max_{b \in \{1, \dots, k\}} \{F_{ib}\}.$$

**Objectives:** For  $x_i \sim \mathcal{N}(\mu_a, C_a)$ , and as  $n, p \rightarrow \infty$ , ( $n_u, n_l \rightarrow \infty$  or  $n_u \rightarrow \infty$ ,  $n_l = O(1)$ )

- Tractable approximation (in norm) for the vectors  $[F_{(u)}]_{\cdot, a}$ ,  $a = 1, \dots, k$
- Joint asymptotic behavior of  $[F_{(u)}]_{i, \cdot}$ .  
⇒ From which classification probability is retrieved.
- Understanding the impact of  $\alpha$   
⇒ Finding optimal  $\alpha$  choice online?

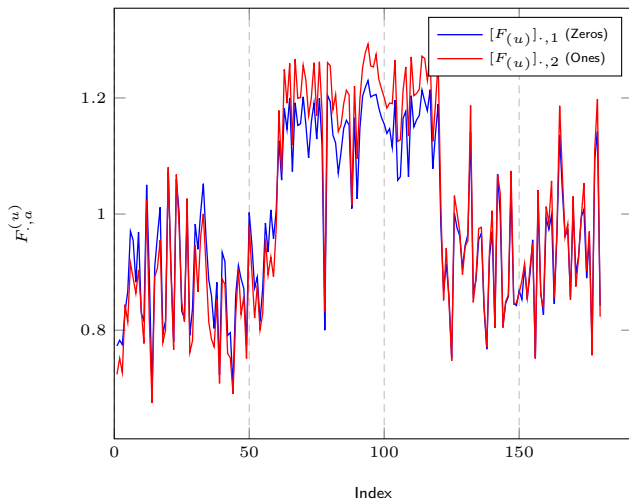


## MNIST Data Example



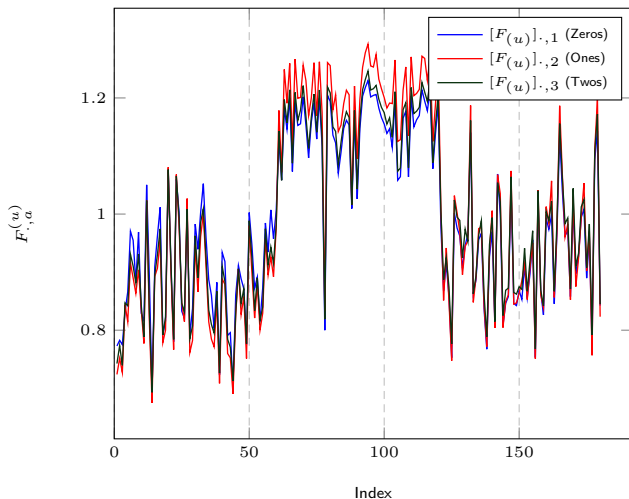
**Figure:** Vectors  $[F^{(u)}]_{.,a}$ ,  $a = 1, 2, 3$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example



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

## MNIST Data Example



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

**Not at all what we expect!:**

**Not at all what we expect!:**

- ▶ Intuitively,  $[F^{(u)}]_{i,a}$  should be close to 1 if  $x_i \in \mathcal{C}_a$  or 0 if  $x_i \notin \mathcal{C}_a$  (from cost function  $K_{ij}(F_{i,a} - F_{j,a})^2$ )

## Not at all what we expect!:

- ▶ Intuitively,  $[F^{(u)}]_{i,a}$  should be close to 1 if  $x_i \in \mathcal{C}_a$  or 0 if  $x_i \notin \mathcal{C}_a$  (from cost function  $K_{ij}(F_{i,a} - F_{j,a})^2$ )
- ▶ Here, strong class-wise biases

**Not at all what we expect!:**

- ▶ Intuitively,  $[F^{(u)}]_{i,a}$  should be close to 1 if  $x_i \in \mathcal{C}_a$  or 0 if  $x_i \notin \mathcal{C}_a$  (from cost function  $K_{ij}(F_{i,a} - F_{j,a})^2$ )
- ▶ Here, **strong class-wise biases**
- ▶ **But, more surprisingly, it still works very well !**

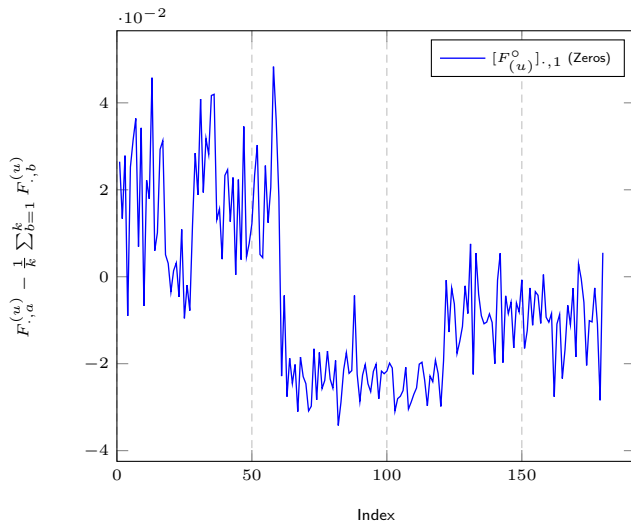
**Not at all what we expect!:**

- ▶ Intuitively,  $[F^{(u)}]_{i,a}$  should be close to 1 if  $x_i \in \mathcal{C}_a$  or 0 if  $x_i \notin \mathcal{C}_a$  (from cost function  $K_{ij}(F_{i,a} - F_{j,a})^2$ )
- ▶ Here, **strong class-wise biases**
- ▶ **But, more surprisingly, it still works very well !**

We need to understand why...

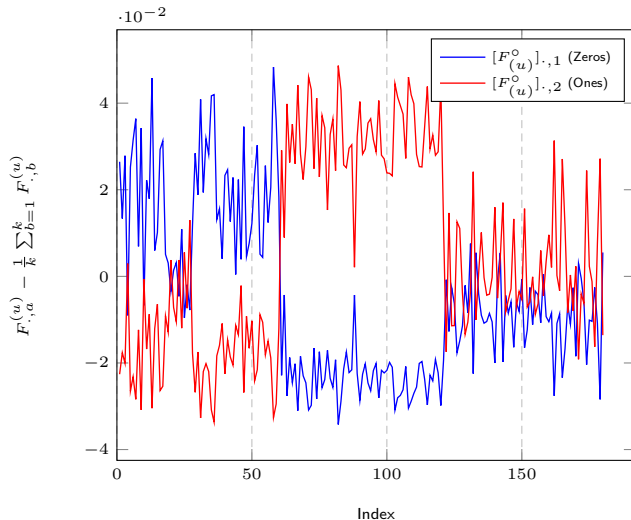


## MNIST Data Example



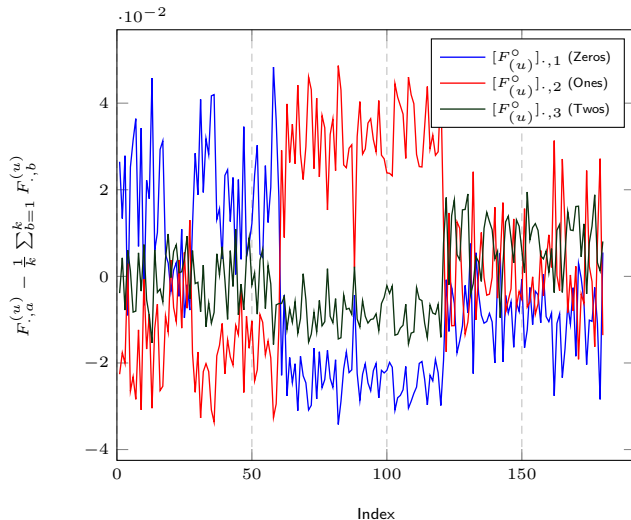
**Figure:** Centered Vectors  $[F^{(u)\circ}_{(u)}]_{\cdot,a} = [F^{(u)}_{(u)} - \frac{1}{k} F^{(u)}_{(u)} \mathbf{1}_k \mathbf{1}_k^T]_{\cdot,a}$ ,  $a = 1, 2, 3$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example



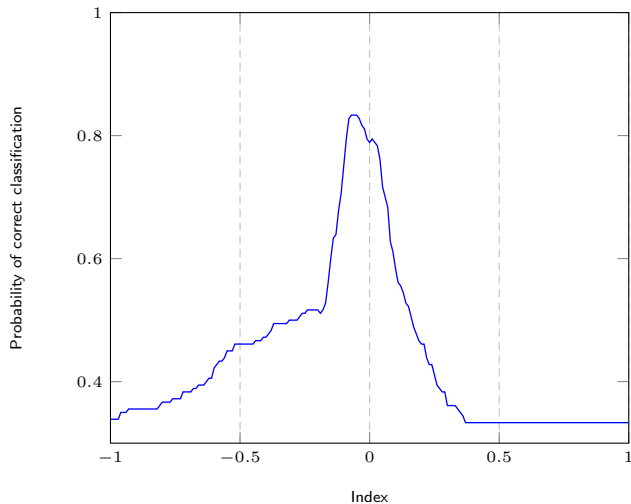
**Figure:** Centered Vectors  $[F^{(u)}_{(u)}]_{.,a} = [F^{(u)}_{(u)} - \frac{1}{k} F^{(u)}_{(u)} 1_k 1_k^T]_{.,a}$ ,  $a = 1, 2, 3$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example



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

## MNIST Data Example



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

# Theoretical Findings

**Method:** We assume  $n_l/n \rightarrow c_l \in (0, 1)$  (“numerous” labelled data setting)

- ▶ Recall that we aim at characterizing

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

- ▶ A priori difficulty linked to **resolvent of involved random matrix!**
- ▶ Painstaking product of complex matrices.

# Theoretical Findings

**Method:** We assume  $n_l/n \rightarrow c_l \in (0, 1)$  (“numerous” labelled data setting)

- ▶ Recall that we aim at characterizing

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

- ▶ A priori difficulty linked to **resolvent of involved random matrix!**
- ▶ Painstaking product of complex matrices.
- ▶ Using Taylor expansion of  $K$  as  $n, p \rightarrow \infty$ , we get

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

$$D_{(u)} = n f(\tau) I_{n_u} + O(n^{\frac{1}{2}})$$

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

# Theoretical Findings

**Method:** We assume  $n_l/n \rightarrow c_l \in (0, 1)$  (“numerous” labelled data setting)

- Recall that we aim at characterizing

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

- A priori difficulty linked to **resolvent of involved random matrix!**
- Painstaking product of complex matrices.
- Using Taylor expansion of  $K$  as  $n, p \rightarrow \infty$ , we get

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

$$D_{(u)} = n f(\tau) I_{n_u} + O(n^{\frac{1}{2}})$$

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

- So that

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

which can be **easily Taylor expanded!**

## Results:

- In the first order,

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

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



## Results:

- In the first order,

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

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

- Many consequences:

## Results:

- In the first order,

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

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

- Many consequences:
  - Random non-informative bias linked to  $v$

## Results:

- In the first order,

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

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

- Many consequences:
  - Random non-informative bias linked to  $v$
  - Strong Impact of  $n_{l,a}$ !
    - ⇒ All  $n_{l,a}$  must be equal OR  $F^{(u)}$  need be scaled!

# Main Results

## Results:

- In the first order,

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

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

- Many consequences:
  - Random non-informative bias linked to  $v$
  - Strong Impact of  $n_{l,a}$ !
    - ⇒ All  $n_{l,a}$  must be equal **OR**  $F^{(u)}$  need be scaled!
  - Additional per-class bias  $\alpha t_a 1_{n_u}$ : no information here
    - ⇒ **Forces the choice**

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

# Main Results

## Results:

- ▶ In the first order,

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

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

- ▶ Many consequences:

- ▶ Random non-informative bias linked to  $v$
- ▶ Strong Impact of  $n_{l,a}$ !
  - ⇒ All  $n_{l,a}$  must be equal **OR**  $F^{(u)}$  need be scaled!
- ▶ Additional per-class bias  $\alpha t_a 1_{n_u}$ : no information here
  - ⇒ **Forces the choice**

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

- ▶ Relevant information hidden in smaller order terms!

# Main Results

As a consequence of the remarks above, we take

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

and define

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

# Main Results

As a consequence of the remarks above, we take

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

and define

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

## Theorem

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

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

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

$$(m_b)_a = -\frac{2f'(\tau)}{f(\tau)} \tilde{M}_{ab} + \frac{f''(\tau)}{f(\tau)} \tilde{t}_a \tilde{t}_b + \frac{2f''(\tau)}{f(\tau)} \tilde{T}_{ab} - \frac{f'(\tau)^2}{f(\tau)^2} t_a t_b + \beta \frac{n}{n_l} \frac{f'(\tau)}{f(\tau)} t_a + B_b$$
$$(\Sigma_b)_{a_1 a_2} = \frac{2\text{tr} C_b^2}{p} \left( \frac{f'(\tau)^2}{f(\tau)^2} + \frac{f''(\tau)}{f(\tau)} \right)^2 t_{a_1} t_{a_2} + \frac{4f'(\tau)^2}{f(\tau)^2} \left( [M^\top C_b M]_{a_1 a_2} + \frac{\delta_{a_1}^2 p}{n_{l,a_1}} T_{ba_1} \right)$$

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  $a \neq b$ ,

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



## Corollary (Asymptotic Classification Error)

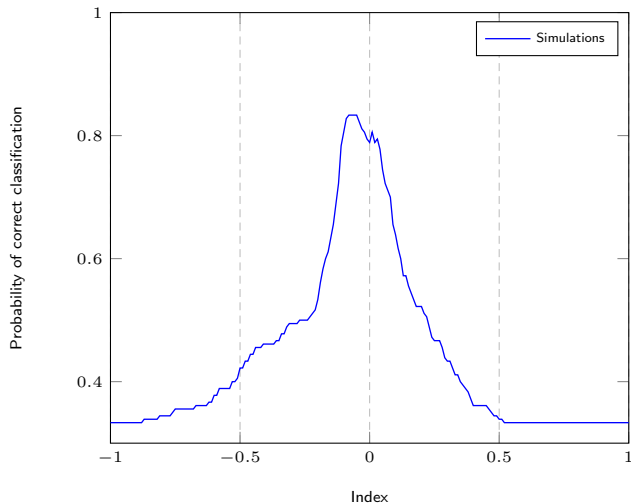
For  $k = 2$  classes and  $a \neq b$ ,

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

**Some consequences:**

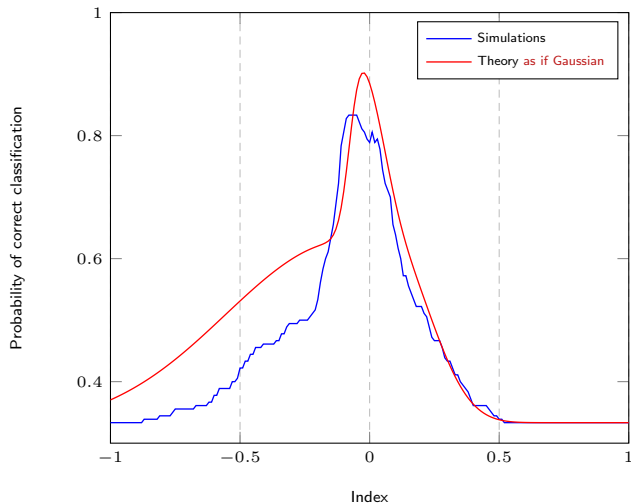
- ▶ non obvious choices of appropriate kernels
- ▶ non obvious choice of optimal  $\beta$  (induces a possibly beneficial bias)
- ▶ importance of  $n_l$  versus  $n_u$ .

## MNIST Data Example



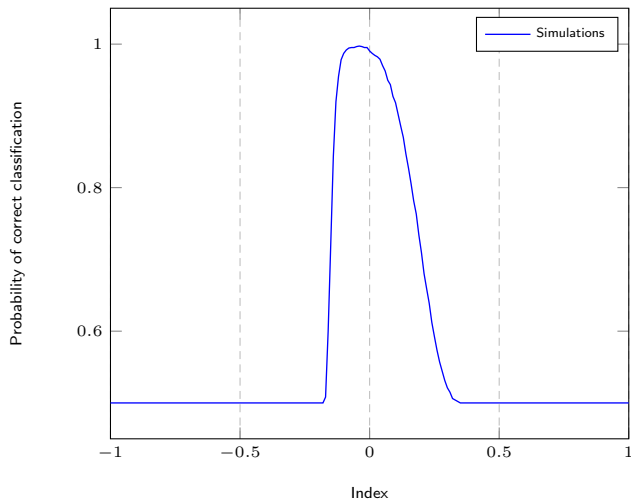
**Figure:** Performance as a function of  $\alpha$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example



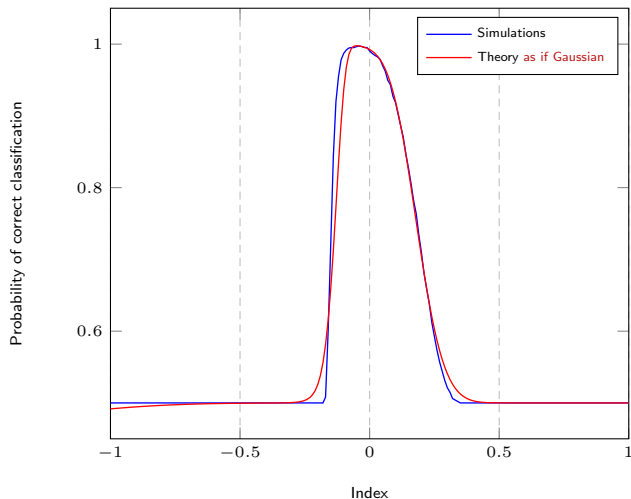
**Figure:** Performance as a function of  $\alpha$ , for 3-class MNIST data (zeros, ones, twos),  $n = 192$ ,  $p = 784$ ,  $n_l/n = 1/16$ , Gaussian kernel.

## MNIST Data Example



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

## MNIST Data Example



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

Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

Kernel Spectral Clustering

Kernel Spectral Clustering: Subspace Clustering

Semi-supervised Learning

**Support Vector Machines**

Neural Networks: Extreme Learning Machines

Random Matrices and Robust Estimation

Perspectives

# Problem Statement

**Context:** All data are labelled, we classify the next incoming one:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k = 2$  classes.

# Problem Statement

**Context:** All data are labelled, we classify the next incoming one:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k = 2$  classes.
- ▶ For kernel  $K(x, y) = \phi(x)^\top \phi(y)$ ,  $\phi(x) \in \mathbb{R}^q$ , find hyperplane directed by  $(w, b)$  to “isolate each class”.

$$(w, b) = \operatorname{argmin}_{w \in \mathbb{R}^{q-1}} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n c(x_i; w, b)$$

for a certain cost function  $c(x; w, b)$ .



# Problem Statement

**Context:** All data are labelled, we classify the next incoming one:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k = 2$  classes.
- ▶ For kernel  $K(x, y) = \phi(x)^\top \phi(y)$ ,  $\phi(x) \in \mathbb{R}^q$ , find hyperplane directed by  $(w, b)$  to “isolate each class”.

$$(w, b) = \operatorname{argmin}_{w \in \mathbb{R}^{q-1}} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n c(x_i; w, b)$$

for a certain cost function  $c(x; w, b)$ .

**Solutions:**

- ▶ **Classical SVM:**

$$c(x_i; w, b) = \mathbb{1}_{\{y_i(w^\top \phi(x_i) + b) \geq 1\}}$$

with  $y_i = \pm 1$  depending on class.

⇒ Solved by **quadratic programming methods**.

⇒ Analysis requires **joint RMT + convex optimization** tools (very interesting but left for later...).

# Problem Statement

**Context:** All data are labelled, we classify the next incoming one:

- ▶ Classify  $x_1, \dots, x_n \in \mathbb{R}^p$  in  $k = 2$  classes.
- ▶ For kernel  $K(x, y) = \phi(x)^\top \phi(y)$ ,  $\phi(x) \in \mathbb{R}^q$ , find hyperplane directed by  $(w, b)$  to “isolate each class”.

$$(w, b) = \operatorname{argmin}_{w \in \mathbb{R}^{q-1}} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n c(x_i; w, b)$$

for a certain cost function  $c(x; w, b)$ .

**Solutions:**

- ▶ **Classical SVM:**

$$c(x_i; w, b) = \mathbb{1}_{\{y_i(w^\top \phi(x_i) + b) \geq 1\}}$$

with  $y_i = \pm 1$  depending on class.

⇒ Solved by **quadratic programming methods**.

⇒ Analysis requires **joint RMT + convex optimization** tools (very interesting but left for later...).

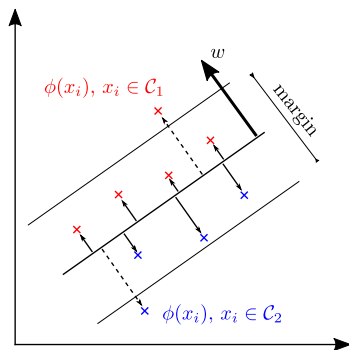
- ▶ **LS SVM:**

$$c(x_i; w, b) = \gamma e_i^2 \equiv \gamma (y_i - w^\top \phi(x_i) - b)^2.$$

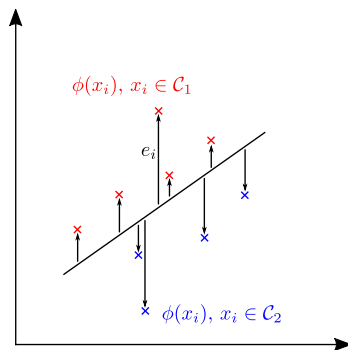
⇒ **Explicit solution** (but not sparse!).

# Problem Statement

Classical SVM



LS SVM



For new datum  $x$ , decision based on (sign of)

$$g(x) = \alpha^\top K(\cdot, x) + b$$

where  $\alpha \in \mathbb{R}^n$  and  $b$  given by

$$\alpha = Q \left( I_n - \frac{1_n 1_n^\top Q}{1_n^\top Q 1_n} \right) y$$
$$b = \frac{1_n^\top Q y}{1_n^\top Q 1_n}$$

where  $Q = (K + \frac{n}{\gamma} I_n)^{-1}$ ,  $y = [y_i]_{i=1}^n$ ,  $\gamma > 0$  some parameter to set.

For new datum  $x$ , decision based on (sign of)

$$g(x) = \alpha^\top K(\cdot, x) + b$$

where  $\alpha \in \mathbb{R}^n$  and  $b$  given by

$$\alpha = Q \left( I_n - \frac{1_n 1_n^\top Q}{1_n^\top Q 1_n} \right) y$$

$$b = \frac{1_n^\top Q y}{1_n^\top Q 1_n}$$

where  $Q = (K + \frac{n}{\gamma} I_n)^{-1}$ ,  $y = [y_i]_{i=1}^n$ ,  $\gamma > 0$  some parameter to set.

### Objectives:

- Study behavior of  $g(x)$

For new datum  $x$ , decision based on (sign of)

$$g(x) = \alpha^\top K(\cdot, x) + b$$

where  $\alpha \in \mathbb{R}^n$  and  $b$  given by

$$\alpha = Q \left( I_n - \frac{1_n 1_n^\top Q}{1_n^\top Q 1_n} \right) y$$

$$b = \frac{1_n^\top Q y}{1_n^\top Q 1_n}$$

where  $Q = (K + \frac{n}{\gamma} I_n)^{-1}$ ,  $y = [y_i]_{i=1}^n$ ,  $\gamma > 0$  some parameter to set.

### Objectives:

- ▶ Study behavior of  $g(x)$
- ▶ For  $x \in \mathcal{C}_a$ , determine probability of success.

For new datum  $x$ , decision based on (sign of)

$$g(x) = \alpha^\top K(\cdot, x) + b$$

where  $\alpha \in \mathbb{R}^n$  and  $b$  given by

$$\alpha = Q \left( I_n - \frac{1_n 1_n^\top Q}{1_n^\top Q 1_n} \right) y$$

$$b = \frac{1_n^\top Q y}{1_n^\top Q 1_n}$$

where  $Q = (K + \frac{n}{\gamma} I_n)^{-1}$ ,  $y = [y_i]_{i=1}^n$ ,  $\gamma > 0$  some parameter to set.

### Objectives:

- ▶ Study behavior of  $g(x)$
- ▶ For  $x \in \mathcal{C}_a$ , determine probability of success.
- ▶ Optimize the parameter  $\gamma$  and the kernel  $K$ .

## Results

As before,  $x_i \sim \mathcal{N}(\mu_a, C_a)$ ,  $a = 1, \dots, k$ , with identical growth conditions, here for  $k = 2$ .



# Results

As before,  $x_i \sim \mathcal{N}(\mu_a, C_a)$ ,  $a = 1, \dots, k$ , with identical growth conditions, here for  $k = 2$ .

**Results:** As  $n, p \rightarrow \infty$ ,

► in the first order

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

# Results

As before,  $x_i \sim \mathcal{N}(\mu_a, C_a)$ ,  $a = 1, \dots, k$ , with identical growth conditions, here for  $k = 2$ .

**Results:** As  $n, p \rightarrow \infty$ ,

► in the first order

$$g(x) = \frac{n_2 - n_1}{n} + \frac{0}{\sqrt{p}} + \underbrace{\frac{G(x)}{p}}_{\text{Relevant terms here!}}$$

► asymptotic Gaussian behavior of  $G(x)$ :

## Theorem

For  $x \in \mathcal{C}_b$ ,  $G(x) - G_b \rightarrow 0$ ,  $G_b \sim \mathcal{N}(m_b, \sigma_b)$ , where

$$m_b = \begin{cases} -2c_2 \cdot c_1 c_2 \gamma \mathcal{D}, & b = 1 \\ +2c_1 \cdot c_1 c_2 \gamma \mathcal{D}, & b = 2 \end{cases}$$

$$\mathcal{D} = -2f'(\tau) \|\mu_2 - \mu_1\|^2 + \frac{f''(\tau)}{p} (\text{tr}(C_2 - C_1))^2 + \frac{2f''(\tau)}{p} \text{tr}((C_2 - C_1)^2)$$

$$\sigma_b = 8\gamma^2 c_1^2 c_2^2 \left[ \frac{(f''(\tau))^2}{p^2} (\text{tr}(C_2 - C_1))^2 \text{tr} C_b^2 + 2(f'(\tau))^2 (\mu_2 - \mu_1)^\top C_b (\mu_2 - \mu_1) \right. \\ \left. + \frac{2(f'(\tau))^2}{n} \left( \frac{\text{tr} C_1 C_b}{c_1} + \frac{\text{tr} C_2 C_b}{c_2} \right) \right]$$

## Consequences:

- ▶ Strong class-size bias  
⇒ Proper threshold must depend on  $n_2 - n_1$ .

## Consequences:

- ▶ Strong class-size bias  
⇒ Proper threshold must depend on  $n_2 - n_1$ .
- ▶ Natural cancellation of  $O(n^{-\frac{1}{2}})$  terms.  
⇒ Similar effect as observed in (properly normalized) kernel spectral clustering.
- ▶ Choice of  $\gamma$  asymptotically irrelevant.

## Consequences:

- ▶ Strong class-size bias  
⇒ Proper threshold must depend on  $n_2 - n_1$ .
- ▶ Natural cancellation of  $O(n^{-\frac{1}{2}})$  terms.  
⇒ Similar effect as observed in (properly normalized) kernel spectral clustering.
- ▶ Choice of  $\gamma$  asymptotically irrelevant.
- ▶ Need to choose  $f'(\tau) < 0$  and  $f''(\tau) > 0$  (not the case for clustering or SSL!)

## Theory and simulations of $g(x)$

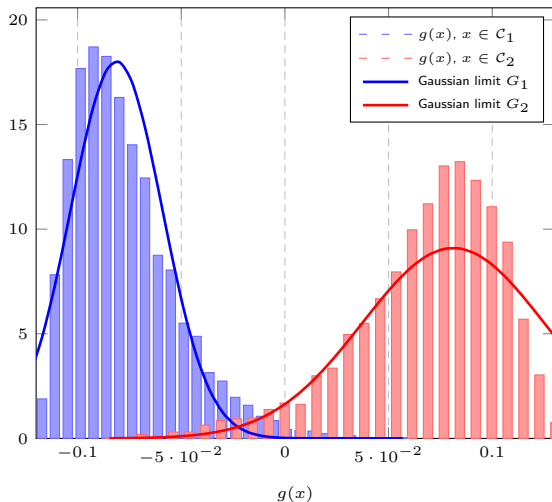
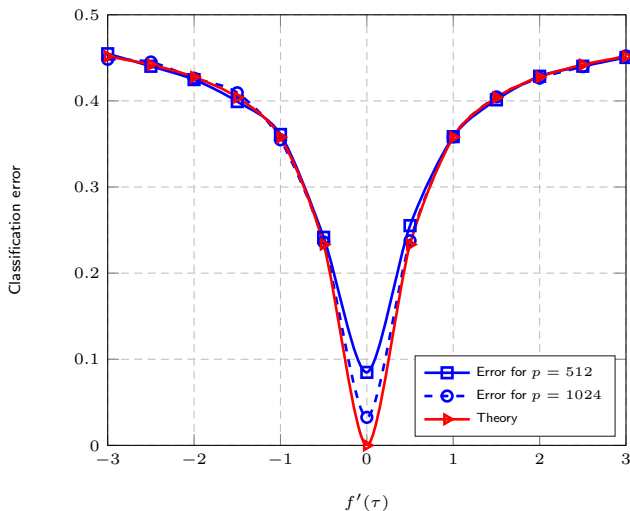


Figure: Values of  $g(x)$  for MNIST data (1's and 7's),  $n = 256$ ,  $p = 784$ , standard Gaussian kernel.

## Classification performance



**Figure:** Performance of LS-SVM,  $c_0 = 1/4$ ,  $c_1 = c_2 = 1/2$ ,  $\gamma = 1$ , polynomial kernel with  $f(\tau) = 4$ ,  $f''(\tau) = 2$ ,  $x \in \mathcal{N}(0, C_a)$ , with  $C_1 = I_p$ ,  $[C_2]_{i,j} = .4^{|i-j|}$ .

Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

Kernel Spectral Clustering

Kernel Spectral Clustering: Subspace Clustering

Semi-supervised Learning

Support Vector Machines

**Neural Networks: Extreme Learning Machines**

Random Matrices and Robust Estimation

Perspectives



## **General plan for the study of neural networks:**

- ▶ Objective is to study performance of neural networks:

## General plan for the study of neural networks:

- ▶ Objective is to study performance of neural networks:
  - ▶ **linear or not** (linear is easy but not interesting, non-linear is hard)

## General plan for the study of neural networks:

- ▶ Objective is to study performance of neural networks:
  - ▶ linear or not (linear is easy but not interesting, non-linear is hard)
  - ▶ from shallow to deep

## General plan for the study of neural networks:

- ▶ Objective is to study performance of neural networks:
  - ▶ linear or not (linear is easy but not interesting, non-linear is hard)
  - ▶ from shallow to deep
  - ▶ recurrent or not (dynamic systems, stability considerations)

## General plan for the study of neural networks:

- ▶ Objective is to study performance of neural networks:
  - ▶ linear or not (linear is easy but not interesting, non-linear is hard)
  - ▶ from shallow to deep
  - ▶ recurrent or not (dynamic systems, stability considerations)
  - ▶ back-propagated or not (LS regression versus gradient descent approaches)

## General plan for the study of neural networks:

- ▶ Objective is to study performance of neural networks:
  - ▶ linear or not (linear is easy but not interesting, non-linear is hard)
  - ▶ from shallow to deep
  - ▶ recurrent or not (dynamic systems, stability considerations)
  - ▶ back-propagated or not (LS regression versus gradient descent approaches)
- ▶ Starting point: simple networks

## General plan for the study of neural networks:

- ▶ Objective is to study performance of neural networks:
  - ▶ **linear or not** (linear is easy but not interesting, non-linear is hard)
  - ▶ **from shallow to deep**
  - ▶ **recurrent or not** (dynamic systems, stability considerations)
  - ▶ **back-propagated or not** (LS regression versus gradient descent approaches)
- ▶ Starting point: simple networks
  - ▶ **Extreme learning machines**: single layer, randomly connected input, LS regressed output.

## General plan for the study of neural networks:

- ▶ Objective is to study performance of neural networks:
  - ▶ **linear or not** (linear is easy but not interesting, non-linear is hard)
  - ▶ **from shallow to deep**
  - ▶ **recurrent or not** (dynamic systems, stability considerations)
  - ▶ **back-propagated or not** (LS regression versus gradient descent approaches)
- ▶ Starting point: simple networks
  - ▶ **Extreme learning machines**: single layer, randomly connected input, LS regressed output.
  - ▶ **Echo-state networks**: single **interconnected** layer, randomly connected input, LS regressed output.



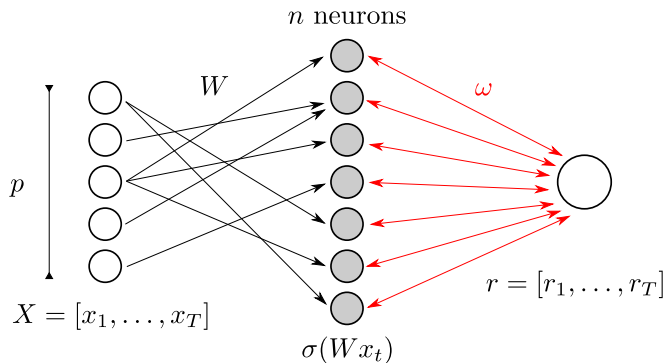
## General plan for the study of neural networks:

- ▶ Objective is to study performance of neural networks:
  - ▶ **linear or not** (linear is easy but not interesting, non-linear is hard)
  - ▶ **from shallow to deep**
  - ▶ **recurrent or not** (dynamic systems, stability considerations)
  - ▶ **back-propagated or not** (LS regression versus gradient descent approaches)
- ▶ Starting point: simple networks
  - ▶ **Extreme learning machines**: single layer, randomly connected input, LS regressed output.
  - ▶ **Echo-state networks**: single **interconnected** layer, randomly connected input, LS regressed output.
  - ▶ **Deeper structures**: back-propagation of error.

# Extreme Learning Machines

**Context:** for a learning period  $T$

- ▶ input vectors  $x_1, \dots, x_T \in \mathbb{R}^p$ , output scalars (or binary values)  $r_1, \dots, r_T \in \mathbb{R}$
- ▶  $n$ -neuron layer, randomly connected input  $W \in \mathbb{R}^{n \times p}$
- ▶ ridge-regressed output  $\omega \in \mathbb{R}^n$
- ▶ non-linear activation function  $\sigma$ .



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

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

- ▶ Training MSE:

$$E_{\gamma}(X, r) = \frac{1}{T} \|r - \omega^{\top} \Sigma\|^2$$

with

$$\Sigma = [\sigma(Wx_1), \dots, \sigma(Wx_T)]$$

$$\omega = \frac{1}{T} \Sigma \left( \frac{1}{T} \Sigma^{\top} \Sigma + \gamma I_T \right)^{-1} r.$$

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

- ▶ Training MSE:

$$E_{\gamma}(X, r) = \frac{1}{T} \|r - \omega^{\top} \Sigma\|^2$$

with

$$\begin{aligned}\Sigma &= [\sigma(Wx_1), \dots, \sigma(Wx_T)] \\ \omega &= \frac{1}{T} \Sigma \left( \frac{1}{T} \Sigma^{\top} \Sigma + \gamma I_T \right)^{-1} r.\end{aligned}$$

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

$$\hat{E}_{\gamma}(X, r; \hat{X}, \hat{r}) = \frac{1}{\hat{T}} \|\hat{r} - \omega^{\top} \sigma(W\hat{X})\|^2.$$

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

- ▶ Training MSE:

$$E_{\gamma}(X, r) = \frac{1}{T} \|r - \omega^{\top} \Sigma\|^2$$

with

$$\begin{aligned}\Sigma &= [\sigma(Wx_1), \dots, \sigma(Wx_T)] \\ \omega &= \frac{1}{T} \Sigma \left( \frac{1}{T} \Sigma^{\top} \Sigma + \gamma I_T \right)^{-1} r.\end{aligned}$$

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

$$\hat{E}_{\gamma}(X, r; \hat{X}, \hat{r}) = \frac{1}{\hat{T}} \|\hat{r} - \omega^{\top} \sigma(W\hat{X})\|^2.$$

- ▶ Optimize over  $\gamma$ .

### Training MSE:

- ▶ Training MSE given by

$$E_{\gamma}(X, r) = \gamma^2 \frac{1}{T} r^{\top} Q_{\gamma}^2 r$$
$$Q_{\gamma} = \left( \frac{1}{T} \Sigma^{\top} \Sigma + \gamma I_T \right)^{-1}.$$

## Training MSE:

- ▶ Training MSE given by

$$E_{\gamma}(X, r) = \gamma^2 \frac{1}{T} r^{\top} Q_{\gamma}^2 r$$
$$Q_{\gamma} = \left( \frac{1}{T} \Sigma^{\top} \Sigma + \gamma I_T \right)^{-1}.$$

- ▶ Testing MSE given by

$$\hat{E}_{\gamma}(X, r; \hat{X}, \hat{r}) = \frac{1}{\hat{T}} \left\| \hat{r} - \frac{1}{\hat{T}} \sigma(W \hat{X})^{\top} \Sigma Q_{\gamma} r \right\|^2$$



## Training MSE:

- ▶ Training MSE given by

$$E_{\gamma}(X, r) = \gamma^2 \frac{1}{T} r^{\top} Q_{\gamma}^2 r$$
$$Q_{\gamma} = \left( \frac{1}{T} \Sigma^{\top} \Sigma + \gamma I_T \right)^{-1}.$$

- ▶ Testing MSE given by

$$\hat{E}_{\gamma}(X, r; \hat{X}, \hat{r}) = \frac{1}{\hat{T}} \left\| \hat{r} - \frac{1}{T} \sigma(W \hat{X})^{\top} \Sigma Q_{\gamma} r \right\|^2$$

- ▶ Requires first a **deterministic equivalent**  $\bar{Q}_{\gamma}$  for  $Q_{\gamma}$  with **non-linear**  $\sigma(\cdot)$ .

### Training MSE:

- ▶ Training MSE given by

$$E_{\gamma}(X, r) = \gamma^2 \frac{1}{T} r^{\top} Q_{\gamma}^2 r$$
$$Q_{\gamma} = \left( \frac{1}{T} \Sigma^{\top} \Sigma + \gamma I_T \right)^{-1}.$$

- ▶ Testing MSE given by

$$\hat{E}_{\gamma}(X, r; \hat{X}, \hat{r}) = \frac{1}{\hat{T}} \left\| \hat{r} - \frac{1}{\hat{T}} \sigma(W \hat{X})^{\top} \Sigma Q_{\gamma} r \right\|^2$$

- ▶ Requires first a **deterministic equivalent**  $\bar{Q}_{\gamma}$  for  $Q_{\gamma}$  with **non-linear**  $\sigma(\cdot)$ .
- ▶ Then **deterministic approximation** of  $\frac{1}{\hat{T}} \sigma(W a)^{\top} \Sigma Q_{\gamma} b$  for deterministic  $a, b$ .

**Main technical difficulty:**  $\Sigma = \sigma(WX) \in \mathbb{R}^{n \times T}$  has

- ▶ independent rows
- ▶ a highly non trivial columns dependence!

**Main technical difficulty:**  $\Sigma = \sigma(WX) \in \mathbb{R}^{n \times T}$  has

- ▶ independent rows
- ▶ a highly non trivial columns dependence!

**Broken trace lemma!:** for  $w \sim \mathcal{N}(0, n^{-1}I_n)$ ,  $X, A$  deterministic of bounded norm,

$$w^\top XAX^\top w \simeq \frac{1}{n} \text{tr} XAX^\top$$

**Main technical difficulty:**  $\Sigma = \sigma(WX) \in \mathbb{R}^{n \times T}$  has

- ▶ independent rows
- ▶ a highly non trivial columns dependence!

**Broken trace lemma!:** for  $w \sim \mathcal{N}(0, n^{-1}I_n)$ ,  $X, A$  deterministic of bounded norm,

$$w^\top X A X^\top w \simeq \frac{1}{n} \text{tr} X A X^\top$$

**BUT** what about:

$$\sigma(w^\top X) A \sigma(X^\top w) \simeq ?$$

## Updated trace lemma:

### Lemma

For  $A$  deterministic and  $\sigma(t)$  polynomial,  $w \in \mathbb{R}^p$  with i.i.d. entries,  $E[w_i] = 0$ ,  $E[w_i^k] = \frac{m_k}{n^{k/2}}$ ,

$$\frac{1}{T} \sigma(w^\top X) A \sigma(X^\top w) - \frac{1}{T} \text{tr} \Phi_X A \xrightarrow{\text{a.s.}} 0$$

with

$$\Phi_X = E \left[ \sigma(X^\top w) \sigma(w^\top X) \right].$$

## Updated trace lemma:

### Lemma

For  $A$  deterministic and  $\sigma(t)$  polynomial,  $w \in \mathbb{R}^p$  with i.i.d. entries,  $E[w_i] = 0$ ,  $E[w_i^k] = \frac{m_k}{n^{k/2}}$ ,

$$\frac{1}{T} \sigma(w^\top X) A \sigma(X^\top w) - \frac{1}{T} \text{tr} \Phi_X A \xrightarrow{\text{a.s.}} 0$$

with

$$\Phi_X = E \left[ \sigma(X^\top w) \sigma(w^\top X) \right].$$

## Technique of proof:

- ▶ Use concentration of vector  $w$
- ▶ transfer concentration by Lipschitz property through mapping  $w \mapsto \sigma(w^\top X)$ , i.e.,

$$P \left( f \left( \sigma(w^\top X) \right) - E \left[ f \left( \sigma(w^\top X) \right) \right] > t \right) \leq c_1 e^{-c_2 n t^2}$$

for all Lipschitz  $f$  (and beyond...), with  $c_1, c_2 > 0$ .

## Results:

- Deterministic equivalent: as  $n, p, T \rightarrow \infty$  with  $\sigma(t)$  smooth,  $W_{ij}$  i.i.d.  
 $E[W_{ij}] = 0$ ,  $E[W_{ij}^k] = \frac{m_k}{n^{k/2}}$ ,

$$Q_\gamma \leftrightarrow \bar{Q}_\gamma$$

where

$$Q_\gamma \left( \frac{1}{T} \Sigma \Sigma^\top + \gamma I_T \right)^{-1}$$
$$\bar{Q}_\gamma = \left( \frac{n}{T} \frac{1}{1 + \delta} \Phi_{\mathbf{X}} + \gamma I_T \right)^{-1}$$

with  $\delta$  unique solution to

$$\delta = \frac{1}{T} \text{tr} \Phi_{\mathbf{X}} \left( \frac{n}{T} \frac{1}{1 + \delta} \Phi_{\mathbf{X}} + \gamma I_T \right)^{-1}.$$



## Neural Network Performances:

- ▶ Training performance:

$$E_{\gamma}(X, r) \leftrightarrow \gamma^2 \frac{1}{T} r^{\top} \bar{Q}_{\gamma} \left[ \frac{\frac{1}{n} \text{tr}(\Psi_X \bar{Q}_{\gamma}^2)}{1 - \frac{1}{n} \text{tr}(\Psi_X \bar{Q}_{\gamma})^2} \Psi_X + I_T \right] \bar{Q}_{\gamma} r.$$

## Neural Network Performances:

- ▶ Training performance:

$$E_{\gamma}(X, r) \leftrightarrow \gamma^2 \frac{1}{T} r^{\top} \bar{Q}_{\gamma} \left[ \frac{\frac{1}{n} \text{tr}(\Psi_X \bar{Q}_{\gamma}^2)}{1 - \frac{1}{n} \text{tr}(\Psi_X \bar{Q}_{\gamma})^2} \Psi_X + I_T \right] \bar{Q}_{\gamma} r.$$

- ▶ Testing performance:

$$\begin{aligned} \hat{E}_{\gamma}(X, r; \hat{X}, \hat{r}) \leftrightarrow & \frac{1}{\hat{T}} \left\| \hat{r} - \Psi_{X, \hat{X}}^{\top} \bar{Q}_{\gamma} r \right\|^2 + \frac{\frac{1}{n} r^{\top} \bar{Q}_{\gamma} \Psi_X \bar{Q}_{\gamma} r}{1 - \frac{1}{n} \text{tr}(\Psi_X \bar{Q}_{\gamma})^2} \\ & \times \left[ \frac{1}{\hat{T}} \text{tr} \Psi_{\hat{X}} - \frac{\gamma}{\hat{T}} \text{tr}(\bar{Q}_{\gamma} \Psi_{X, \hat{X}} \Psi_{\hat{X}, X} \bar{Q}_{\gamma}) - \frac{1}{\hat{T}} \text{tr}(\Psi_{\hat{X}, X} \bar{Q}_{\gamma}) \Psi_{X, \hat{X}} \right]. \end{aligned}$$

where  $\Psi_{A, B} = \frac{n}{T} \frac{1}{1+\delta} \Phi_{A, B}$ ,  $\Psi_A = \Psi_{A, A}$ ,  $\Phi_{A, B} = E[\frac{1}{n} \sigma(WA)^{\top} \sigma(WB)]$ .

## Neural Network Performances:

- ▶ Training performance:

$$E_{\gamma}(X, r) \leftrightarrow \gamma^2 \frac{1}{T} r^{\top} \bar{Q}_{\gamma} \left[ \frac{\frac{1}{n} \text{tr}(\Psi_X \bar{Q}_{\gamma}^2)}{1 - \frac{1}{n} \text{tr}(\Psi_X \bar{Q}_{\gamma})^2} \Psi_X + I_T \right] \bar{Q}_{\gamma} r.$$

- ▶ Testing performance:

$$\begin{aligned} \hat{E}_{\gamma}(X, r; \hat{X}, \hat{r}) \leftrightarrow & \frac{1}{\hat{T}} \left\| \hat{r} - \Psi_{X, \hat{X}}^{\top} \bar{Q}_{\gamma} r \right\|^2 + \frac{\frac{1}{n} r^{\top} \bar{Q}_{\gamma} \Psi_X \bar{Q}_{\gamma} r}{1 - \frac{1}{n} \text{tr}(\Psi_X \bar{Q}_{\gamma})^2} \\ & \times \left[ \frac{1}{\hat{T}} \text{tr} \Psi_{\hat{X}} - \frac{\gamma}{\hat{T}} \text{tr}(\bar{Q}_{\gamma} \Psi_{X, \hat{X}} \Psi_{\hat{X}, X} \bar{Q}_{\gamma}) - \frac{1}{\hat{T}} \text{tr}(\Psi_{\hat{X}, X} \bar{Q}_{\gamma}) \Psi_{X, \hat{X}} \right]. \end{aligned}$$

where  $\Psi_{A,B} = \frac{n}{T} \frac{1}{1+\delta} \Phi_{A,B}$ ,  $\Psi_A = \Psi_{A,A}$ ,  $\Phi_{A,B} = E[\frac{1}{n} \sigma(WA)^{\top} \sigma(WB)]$ .

**In the limit where  $n/p, n/T \rightarrow \infty$ , taking  $\gamma = \frac{n}{T} \Gamma$ :**

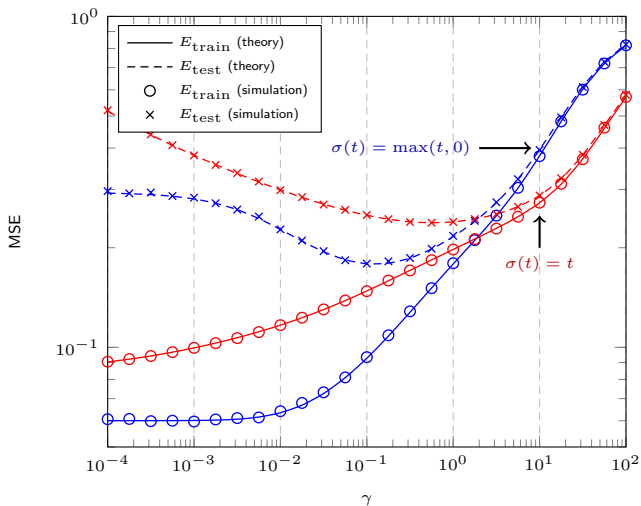
$$\begin{aligned} E_{\gamma}(X, r) & \leftrightarrow \frac{1}{T} \Gamma^2 r^{\top} (\Phi_X + \Gamma I_T)^{-2} r \\ \hat{E}_{\gamma}(X, r) & \leftrightarrow \frac{1}{\hat{T}} \left\| \hat{r} - \Phi_{\hat{X}, X} (\Phi_X + \Gamma I_T)^{-1} r \right\|^2. \end{aligned}$$

## Special Cases of $\Phi_{A,B}$ :

$\sigma(t)$	$W_{ij}$	$[\Phi_{A,B}]_{ij}$
$t$	any	$\frac{m_2}{n} a_i^\top b_j$
$At^2 + Bt + C$	any	$A^2 \left[ \frac{m_2^2}{n^2} \left( 2(a_i^\top b_j)^2 + \ a_i\ ^2 \ b_j\ ^2 \right) + \frac{m_4 - 3m_2^2}{n^2} (a_i^2)^\top (b_j^2) \right]$ $+ B^2 \frac{m_2}{n} a_i^\top b_j + AB \frac{m_3}{n^{3/2}} \left[ (a_i^2)^\top b_j + a_i^\top (b_j^2) \right]$ $+ AC \frac{m_2}{n} [\ a_i\ ^2 + \ b_j\ ^2] + C^2$
$\max(t, 0)$	$\mathcal{N}(0, \frac{1}{n})$	$\frac{1}{2\pi n} \ a_i\  \ b_j\  \left( Z_{ij} (-Z_{ij}) + \sqrt{1 - Z_{ij}^2} \right)$
$\text{erf}(t)$	$\mathcal{N}(0, \frac{1}{n})$	$\frac{2}{\pi} \left( \frac{2a_i^\top b_j}{\sqrt{(n+2\ a_i\ ^2)(n+2\ b_j\ ^2)}} \right)$
$1_{\{t>0\}}$	$\mathcal{N}(0, \frac{1}{n})$	$\frac{1}{2} - \frac{1}{2\pi} (Z_{ij})$
$\text{sign}(t)$	$\mathcal{N}(0, \frac{1}{n})$	$1 - \frac{1}{\pi} (Z_{ij})$
$\cos(t)$	$\mathcal{N}(0, \frac{1}{n})$	$\exp \left( -\frac{1}{2} [\ a_i\ ^2 + \ b_j\ ^2] \right) \cosh \left( a_i^\top b_j \right).$

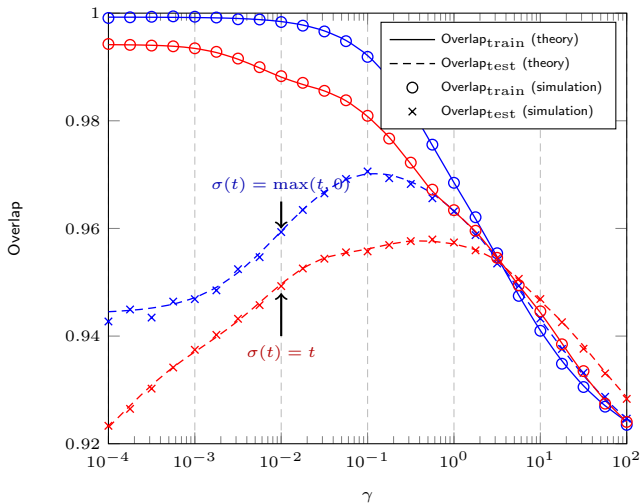
**Figure:**  $\Phi_{A,B}$  for  $W_{ij}$  i.i.d. zero mean,  $k$ -th order moments  $m_k n^{-\frac{k}{2}}$ ,  $Z_{ij} \equiv \frac{a_i^\top b_j}{\|a_i\| \|b_j\|}$ ,  $(a^2) = [a_i^2]_{i=1}^n$ .

## Test on MNIST data



**Figure:** MSE performance for  $\sigma(t) = t$  and  $\sigma(t) = \max(t, 0)$ , as a function of  $\gamma$ , for 2-class MNIST data (sevens, nines),  $n = 512$ ,  $T = 1024$ ,  $p = 784$ .

## Test on MNIST data



**Figure:** Overlap performance for  $\sigma(t) = t$  and  $\sigma(t) = \max(t, 0)$ , as a function of  $\gamma$ , for 2-class MNIST data (sevens, nines),  $n = 512$ ,  $T = 1024$ ,  $p = 784$ .

### Interpretations and Improvements:

- ▶ General formulas for  $\Phi_X$ ,  $\Phi_{X,\hat{x}}$
- ▶ On-line optimization of  $\gamma$ ,  $\sigma(\cdot)$ ,  $n$ ?

## Interpretations and Improvements:

- ▶ General formulas for  $\Phi_X$ ,  $\Phi_{X,\hat{x}}$
- ▶ On-line optimization of  $\gamma$ ,  $\sigma(\cdot)$ ,  $n$ ?

## Generalizations:

- ▶ Multi-layer ELM?
- ▶ Optimize layers vs. number of neurons?
- ▶ Backpropagation error analysis?
- ▶ Connection to auto-encoders?
- ▶ Introduction of non-linearity to more involved structures (ESN, deep nets?).



Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

Kernel Spectral Clustering

Kernel Spectral Clustering: Subspace Clustering

Semi-supervised Learning

Support Vector Machines

Neural Networks: Extreme Learning Machines

**Random Matrices and Robust Estimation**

Perspectives

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

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

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

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n x_i x_i^*.$$

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

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

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n x_i x_i^*.$$

- **[Huber'67]** If  $x_1 \sim (1 - \varepsilon)\mathcal{N}(0, C_N) + \varepsilon G$ ,  $G$  unknown, robust estimator ( $n > N$ )

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n \max \left\{ \ell_1, \frac{\ell_2}{\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i} \right\} x_i x_i^* \text{ for some } \ell_1, \ell_2 > 0.$$

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

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

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n x_i x_i^*.$$

- ▶ **[Huber'67]** If  $x_1 \sim (1 - \varepsilon)\mathcal{N}(0, C_N) + \varepsilon G$ ,  $G$  unknown, robust estimator ( $n > N$ )

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n \max \left\{ \ell_1, \frac{\ell_2}{\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i} \right\} x_i x_i^* \text{ for some } \ell_1, \ell_2 > 0.$$

- ▶ **[Maronna'76]** If  $x_1$  elliptical (and  $n > N$ ), ML estimator for  $C_N$  given by

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n u \left( \frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^* \text{ for some non-increasing } u.$$

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

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

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n x_i x_i^*.$$

- **[Huber'67]** If  $x_1 \sim (1 - \varepsilon)\mathcal{N}(0, C_N) + \varepsilon G$ ,  $G$  unknown, robust estimator ( $n > N$ )

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n \max \left\{ \ell_1, \frac{\ell_2}{\frac{1}{N} x_i^* \hat{C}_N^{-1} x_i} \right\} x_i x_i^* \text{ for some } \ell_1, \ell_2 > 0.$$

- **[Maronna'76]** If  $x_1$  elliptical (and  $n > N$ ), ML estimator for  $C_N$  given by

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n u \left( \frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^* \text{ for some non-increasing } u.$$

- **[Pascal'13; Chen'11]** If  $N > n$ ,  $x_1$  elliptical or with outliers, shrinkage extensions

$$\hat{C}_N(\rho) = (1 - \rho) \frac{1}{n} \sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N} x_i^* \hat{C}_N^{-1}(\rho) x_i} + \rho I_N$$

$$\check{C}_N(\rho) = \frac{\check{B}_N(\rho)}{\frac{1}{N} \text{tr } \check{B}_N(\rho)}, \quad \check{B}_N(\rho) = (1 - \rho) \frac{1}{n} \sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N} x_i^* \check{C}_N^{-1}(\rho) x_i} + \rho I_N$$

Results only known for  $N$  fixed and  $n \rightarrow \infty$ :

- ▶ not appropriate in settings of interest today (BigData, array processing, MIMO)

Results only known for  $N$  fixed and  $n \rightarrow \infty$ :

- ▶ not appropriate in settings of interest today (BigData, array processing, MIMO)

We study such  $\hat{C}_N$  in the regime

$$N, n \rightarrow \infty, \quad N/n \rightarrow c \in (0, \infty).$$



Results only known for  $N$  fixed and  $n \rightarrow \infty$ :

- ▶ not appropriate in settings of interest today (BigData, array processing, MIMO)

We study such  $\hat{C}_N$  in the regime

$$N, n \rightarrow \infty, \quad N/n \rightarrow c \in (0, \infty).$$

- ▶ Math interest:
  - ▶ limiting eigenvalue distribution of  $\hat{C}_N$
  - ▶ limiting values and fluctuations of functionals  $f(\hat{C}_N)$

Results only known for  $N$  fixed and  $n \rightarrow \infty$ :

- ▶ not appropriate in settings of interest today (BigData, array processing, MIMO)

We study such  $\hat{C}_N$  in the regime

$$N, n \rightarrow \infty, \quad N/n \rightarrow c \in (0, \infty).$$

- ▶ Math interest:
  - ▶ limiting eigenvalue distribution of  $\hat{C}_N$
  - ▶ limiting values and fluctuations of functionals  $f(\hat{C}_N)$
- ▶ Application interest:
  - ▶ comparison between SCM and robust estimators
  - ▶ performance of robust/non-robust estimation methods
  - ▶ improvement thereof (by proper parametrization)

## Definition (Maronna's Estimator)

For  $x_1, \dots, x_n \in \mathbb{C}^N$  with  $n > N$ ,  $\hat{C}_N$  is the solution (upon existence and uniqueness) of

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n u \left( \frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*$$

## Definition (Maronna's Estimator)

For  $x_1, \dots, x_n \in \mathbb{C}^N$  with  $n > N$ ,  $\hat{C}_N$  is the solution (upon existence and uniqueness) of

$$\hat{C}_N = \frac{1}{n} \sum_{i=1}^n u \left( \frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*$$

where  $u : [0, \infty) \rightarrow (0, \infty)$  is

- ▶ non-increasing
- ▶ such that  $\phi(x) \triangleq xu(x)$  increasing of supremum  $\phi_\infty$  with

$$1 < \phi_\infty < c^{-1}, \quad c \in (0, 1).$$

# The Results in a Nutshell

For various models of the  $x_i$ 's,

- First order convergence:

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

for some **tractable** random matrices  $\hat{S}_N$ .

⇒ We only discuss this result [here](#).

# The Results in a Nutshell

For various models of the  $x_i$ 's,

- First order convergence:

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

for some **tractable** random matrices  $\hat{S}_N$ .

⇒ We only discuss this result [here](#).

- Second order results:

$$N^{1-\varepsilon} \left( a^* \hat{C}_N^k b - a^* \hat{S}_N^k b \right) \xrightarrow{\text{a.s.}} 0$$

allowing **transfer of CLT results**.

# The Results in a Nutshell

For various models of the  $x_i$ 's,

- ▶ First order convergence:

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

for some **tractable** random matrices  $\hat{S}_N$ .

⇒ We only discuss this result here.

- ▶ Second order results:

$$N^{1-\varepsilon} \left( a^* \hat{C}_N^k b - a^* \hat{S}_N^k b \right) \xrightarrow{\text{a.s.}} 0$$

allowing **transfer of CLT results**.

- ▶ Applications:

- ▶ improved robust covariance matrix estimation
- ▶ improved robust tests / estimators
- ▶ specific examples in **statistics** at large, **array processing**, statistical **finance**, etc.

## (Elliptical) scenario

### Theorem (Large dimensional behavior, elliptical case)

For  $x_i = \sqrt{\tau_i} w_i$ ,  $\tau_i$  impulsive (random or not),  $w_i$  unitarily invariant,  $\|w_i\| = N$ ,

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

with, for some  $v$  related to  $u$  ( $v = u \circ g^{-1}$ ,  $g(x) = x(1 - c\phi(x))^{-1}$ ),

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^n u \left( \frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*, \quad \hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^n v(\tau_i \gamma_N) x_i x_i^*$$

and  $\gamma_N$  unique solution of

$$1 = \frac{1}{n} \sum_{j=1}^n \frac{\gamma v(\tau_j \gamma)}{1 + c\gamma v(\tau_j \gamma)}.$$



## (Elliptical) scenario

### Theorem (Large dimensional behavior, elliptical case)

For  $x_i = \sqrt{\tau_i} w_i$ ,  $\tau_i$  impulsive (random or not),  $w_i$  unitarily invariant,  $\|w_i\| = N$ ,

$$\|\hat{C}_N - \hat{S}_N\| \xrightarrow{\text{a.s.}} 0$$

with, for some  $v$  related to  $u$  ( $v = u \circ g^{-1}$ ,  $g(x) = x(1 - c\phi(x))^{-1}$ ),

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^n u \left( \frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*, \quad \hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^n v(\tau_i \gamma_N) x_i x_i^*$$

and  $\gamma_N$  unique solution of

$$1 = \frac{1}{n} \sum_{j=1}^n \frac{\gamma v(\tau_j \gamma)}{1 + c\gamma v(\tau_j \gamma)}.$$

### Corollaries

- **Spectral measure:**  $\mu_{\hat{C}_N} - \mu_{\hat{S}_N} \xrightarrow{\mathcal{L}} 0$  a.s. ( $\mu_N^X \triangleq \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(X)}$ )

## (Elliptical) scenario

### Theorem (Large dimensional behavior, elliptical case)

For  $x_i = \sqrt{\tau_i} w_i$ ,  $\tau_i$  impulsive (random or not),  $w_i$  unitarily invariant,  $\|w_i\| = N$ ,

$$\|\hat{C}_N - \hat{S}_N\| \xrightarrow{\text{a.s.}} 0$$

with, for some  $v$  related to  $u$  ( $v = u \circ g^{-1}$ ,  $g(x) = x(1 - c\phi(x))^{-1}$ ),

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^n u \left( \frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*, \quad \hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^n v(\tau_i \gamma_N) x_i x_i^*$$

and  $\gamma_N$  unique solution of

$$1 = \frac{1}{n} \sum_{j=1}^n \frac{\gamma v(\tau_j \gamma)}{1 + c\gamma v(\tau_j \gamma)}.$$

### Corollaries

- ▶ **Spectral measure:**  $\mu_{\hat{C}_N}^X - \mu_{\hat{S}_N}^X \xrightarrow{\mathcal{L}} 0$  a.s. ( $\mu_N^X \triangleq \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(X)}$ )
- ▶ **Local convergence:**  $\max_{1 \leq i \leq N} |\lambda_i(\hat{C}_N) - \lambda_i(\hat{S}_N)| \xrightarrow{\text{a.s.}} 0$ .

## (Elliptical) scenario

### Theorem (Large dimensional behavior, elliptical case)

For  $x_i = \sqrt{\tau_i} w_i$ ,  $\tau_i$  impulsive (random or not),  $w_i$  unitarily invariant,  $\|w_i\| = N$ ,

$$\|\hat{C}_N - \hat{S}_N\| \xrightarrow{\text{a.s.}} 0$$

with, for some  $v$  related to  $u$  ( $v = u \circ g^{-1}$ ,  $g(x) = x(1 - c\phi(x))^{-1}$ ),

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^n u \left( \frac{1}{N} x_i^* \hat{C}_N^{-1} x_i \right) x_i x_i^*, \quad \hat{S}_N \triangleq \frac{1}{n} \sum_{i=1}^n v(\tau_i \gamma_N) x_i x_i^*$$

and  $\gamma_N$  unique solution of

$$1 = \frac{1}{n} \sum_{j=1}^n \frac{\gamma v(\tau_j \gamma)}{1 + c\gamma v(\tau_j \gamma)}.$$

### Corollaries

- ▶ **Spectral measure:**  $\mu_{\hat{C}_N}^X - \mu_{\hat{S}_N}^X \xrightarrow{\mathcal{L}} 0$  a.s. ( $\mu_N^X \triangleq \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(X)}$ )
- ▶ **Local convergence:**  $\max_{1 \leq i \leq N} |\lambda_i(\hat{C}_N) - \lambda_i(\hat{S}_N)| \xrightarrow{\text{a.s.}} 0$ .
- ▶ **Norm boundedness:**  $\limsup_N \|\hat{C}_N\| < \infty$

→ Bounded spectrum (unlike SCM!)

## Large dimensional behavior

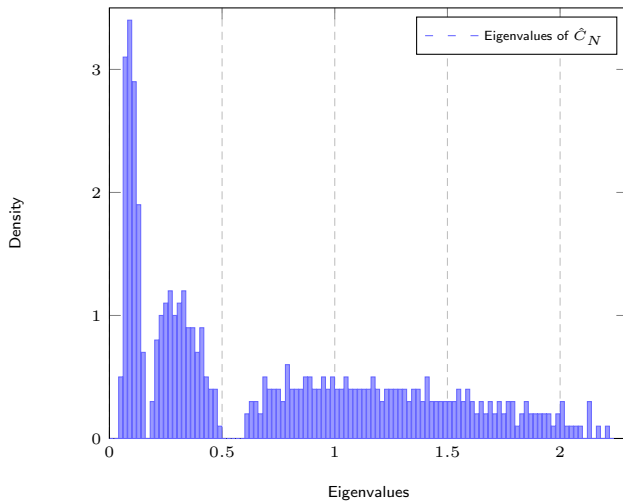


Figure:  $n = 2500$ ,  $N = 500$ ,  $C_N = \text{diag}(I_{125}, 3I_{125}, 10I_{250})$ ,  $\tau_i \sim \Gamma(.5, 2)$  i.i.d.

## Large dimensional behavior

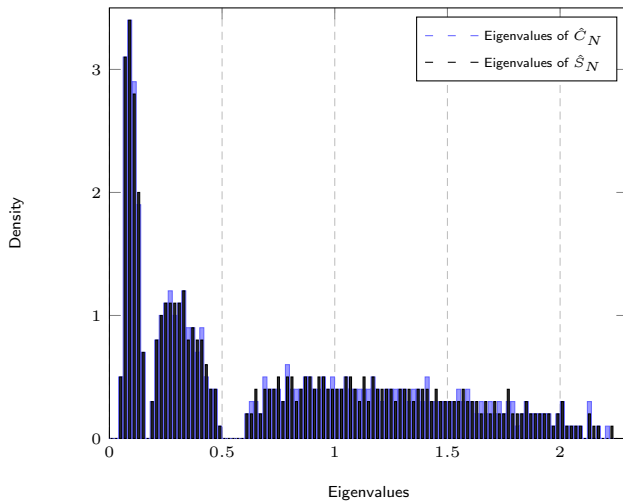


Figure:  $n = 2500$ ,  $N = 500$ ,  $C_N = \text{diag}(I_{125}, 3I_{125}, 10I_{250})$ ,  $\tau_i \sim \Gamma(.5, 2)$  i.i.d.

## Large dimensional behavior

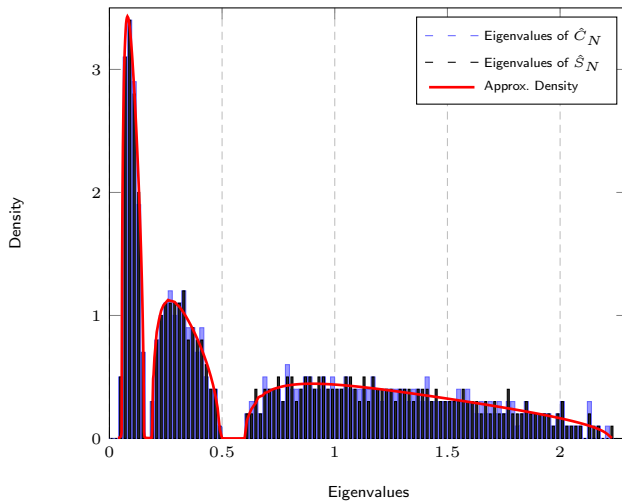


Figure:  $n = 2500$ ,  $N = 500$ ,  $C_N = \text{diag}(I_{125}, 3I_{125}, 10I_{250})$ ,  $\tau_i \sim \Gamma(.5, 2)$  i.i.d.

## Definition ( $v$ and $\psi$ )

Letting  $g(x) = x(1 - c\phi(x))^{-1}$  (on  $\mathbb{R}_+$ ),

$$v(x) \triangleq (u \circ g^{-1})(x) \quad \text{non-increasing}$$

$$\psi(x) \triangleq xv(x) \quad \text{increasing and bounded by } \psi_\infty.$$

## Definition ( $v$ and $\psi$ )

Letting  $g(x) = x(1 - c\phi(x))^{-1}$  (on  $\mathbb{R}_+$ ),

$$v(x) \triangleq (u \circ g^{-1})(x) \quad \text{non-increasing}$$

$$\psi(x) \triangleq xv(x) \quad \text{increasing and bounded by } \psi_\infty.$$

## Lemma (Rewriting $\hat{C}_N$ )

It holds (with  $C_N = I_N$ ) that

$$\hat{C}_N \triangleq \frac{1}{n} \sum_{i=1}^n \tau_i v(\tau_i d_i) w_i w_i^*$$

with  $(d_1, \dots, d_n) \in \mathbb{R}_+^n$  a.s. unique solution to

$$d_i = \frac{1}{N} w_i^* \hat{C}_{(i)}^{-1} w_i = \frac{1}{N} w_i^* \left( \frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} w_i, \quad i = 1, \dots, n.$$



### Remark (Quadratic Form close to Trace)

Random matrix insight:  $(\frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^*)^{-1}$  “almost independent” of  $w_i$ , so

## Remark (Quadratic Form close to Trace)

Random matrix insight:  $(\frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^*)^{-1}$  “almost independent” of  $w_i$ , so

$$d_i = \frac{1}{N} w_i^* \left( \frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} w_i \simeq \frac{1}{N} \text{tr} \left( \frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} \simeq \gamma_N$$

for some deterministic sequence  $(\gamma_N)_{N=1}^{\infty}$ , irrespective of  $i$ .

## Remark (Quadratic Form close to Trace)

Random matrix insight:  $(\frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^*)^{-1}$  “almost independent” of  $w_i$ , so

$$d_i = \frac{1}{N} w_i^* \left( \frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} w_i \simeq \frac{1}{N} \text{tr} \left( \frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j d_j) w_j w_j^* \right)^{-1} \simeq \gamma_N$$

for some deterministic sequence  $(\gamma_N)_{N=1}^\infty$ , irrespective of  $i$ .

## Lemma (Key Lemma)

Letting  $e_i \triangleq \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$  with  $\gamma_N$  unique solution to

$$1 = \frac{1}{n} \sum_{k=1}^n \frac{\psi(\tau_k \gamma_N)}{1 + c\psi(\tau_k \gamma_N)}$$

we have

$$\max_{1 \leq i \leq n} |e_i - 1| \xrightarrow{\text{a.s.}} 0.$$

Proof of the Key Lemma:  $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$ ,  $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Property (Quadratic form and  $\gamma_N$ )

$$\max_{1 \leq i \leq n} \left| \frac{1}{N} w_i^* \left( \frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j \gamma_N) w_j w_j^* \right)^{-1} w_i - \gamma_N \right| \xrightarrow{\text{a.s.}} 0.$$

Proof of the Key Lemma:  $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$ ,  $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Property (Quadratic form and  $\gamma_N$ )

$$\max_{1 \leq i \leq n} \left| \frac{1}{N} w_i^* \left( \frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j \gamma_N) w_j w_j^* \right)^{-1} w_i - \gamma_N \right| \xrightarrow{\text{a.s.}} 0.$$

Proof of the Property

- Uniformity easy (moments of all orders for  $[w_i]_j$ ).
- By a “quadratic form similar to trace” approach, we get

$$\max_{1 \leq i \leq n} \left| \frac{1}{N} w_i^* \left( \frac{1}{n} \sum_{j \neq i} \tau_j v(\tau_j \gamma_N) w_j w_j^* \right)^{-1} w_i - m(0) \right| \xrightarrow{\text{a.s.}} 0$$

with  $m(0)$  unique positive solution to **[MarPas'67; BaiSil'95]**

$$m(0) = \frac{1}{n} \sum_{i=1}^n \frac{\tau_i v(\tau_i \gamma_N)}{1 + c \tau_i v(\tau_i \gamma_N) m(0)}.$$

- $\gamma_N$  precisely solves this equation, thus  $m(0) = \gamma_N$ .

Proof of the Key Lemma:  $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$ ,  $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Substitution Trick (case  $\tau_i \in [a, b] \subset (0, \infty)$ )

Up to relabelling  $e_1 \leq \dots \leq e_n$ , use

$$\begin{aligned} v(\tau_n \gamma_N) \mathbf{e}_n &= v(\tau_n d_n) = v \left( \tau_n \frac{1}{N} w_n^* \left( \frac{1}{n} \sum_{i < n} \tau_i \underbrace{v(\tau_i d_i)}_{=v(\tau_i \gamma_N) \mathbf{e}_i} w_i w_i^* \right)^{-1} w_n \right) \\ &\leq v \left( \tau_n \mathbf{e}_n^{-1} \frac{1}{N} w_n^* \left( \frac{1}{n} \sum_{i < n} \tau_i v(\tau_i \gamma_N) w_i w_i^* \right)^{-1} w_n \right) \\ &\leq v \left( \tau_n \mathbf{e}_n^{-1} (\gamma_N - \varepsilon_n) \right) \text{ a.s., } \varepsilon_n \rightarrow 0 \text{ (slow).} \end{aligned}$$

Proof of the Key Lemma:  $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$ ,  $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

Substitution Trick (case  $\tau_i \in [a, b] \subset (0, \infty)$ )

Up to relabelling  $e_1 \leq \dots \leq e_n$ , use

$$\begin{aligned} v(\tau_n \gamma_N) \mathbf{e}_n &= v(\tau_n d_n) = v \left( \tau_n \frac{1}{N} w_n^* \left( \frac{1}{n} \sum_{i < n} \tau_i \underbrace{v(\tau_i d_i)}_{=v(\tau_i \gamma_N) \mathbf{e}_i} w_i w_i^* \right)^{-1} w_n \right) \\ &\leq v \left( \tau_n \mathbf{e}_n^{-1} \frac{1}{N} w_n^* \left( \frac{1}{n} \sum_{i < n} \tau_i v(\tau_i \gamma_N) w_i w_i^* \right)^{-1} w_n \right) \\ &\leq v \left( \tau_n \mathbf{e}_n^{-1} (\gamma_N - \varepsilon_n) \right) \text{ a.s., } \varepsilon_n \rightarrow 0 \text{ (slow).} \end{aligned}$$

Use properties of  $\psi$  to get

$$\psi(\tau_n \gamma_N) \leq \psi(\tau_n \mathbf{e}_n^{-1} \gamma_N) \left(1 - \varepsilon_n \gamma_N^{-1}\right)^{-1}$$

**Proof of the Key Lemma:**  $\max_i |e_i - 1| \xrightarrow{\text{a.s.}} 0$ ,  $e_i = \frac{v(\tau_i d_i)}{v(\tau_i \gamma_N)}$

**Substitution Trick** (case  $\tau_i \in [a, b] \subset (0, \infty)$ )

Up to relabelling  $e_1 \leq \dots \leq e_n$ , use

$$\begin{aligned} v(\tau_n \gamma_N) \mathbf{e}_n &= v(\tau_n d_n) = v \left( \tau_n \frac{1}{N} w_n^* \left( \frac{1}{n} \sum_{i < n} \tau_i \underbrace{v(\tau_i d_i)}_{=v(\tau_i \gamma_N) \mathbf{e}_i} w_i w_i^* \right)^{-1} w_n \right) \\ &\leq v \left( \tau_n \mathbf{e}_n^{-1} \frac{1}{N} w_n^* \left( \frac{1}{n} \sum_{i < n} \tau_i v(\tau_i \gamma_N) w_i w_i^* \right)^{-1} w_n \right) \\ &\leq v(\tau_n \mathbf{e}_n^{-1} (\gamma_N - \varepsilon_n)) \quad \text{a.s., } \varepsilon_n \rightarrow 0 \text{ (slow).} \end{aligned}$$

Use properties of  $\psi$  to get

$$\psi(\tau_n \gamma_N) \leq \psi(\tau_n \mathbf{e}_n^{-1} \gamma_N) \left(1 - \varepsilon_n \gamma_N^{-1}\right)^{-1}$$

**Conclusion:** If  $e_n > 1 + \ell$  i.o., as  $\tau_n \in [a, b]$ , on subsequence  $\begin{cases} \tau_n \rightarrow \tau_0 > 0 \\ \gamma_N \rightarrow \gamma_0 > 0 \end{cases}$ ,

$$\psi(\tau_0 \gamma_0) \leq \psi\left(\frac{\tau_0 \gamma_0}{1 + \ell}\right), \text{ a contradiction.}$$



## Theorem (Outlier Rejection)

Observation set

$$X = [x_1, \dots, x_{(1-\varepsilon_n)n}, a_1, \dots, a_{\varepsilon_n n}]$$

where  $x_i \sim \mathcal{CN}(0, C_N)$  and  $a_1, \dots, a_{\varepsilon_n n} \in \mathbb{C}^N$  *deterministic* outliers. Then,

$$\|\hat{C}_N - \hat{S}_N\| \xrightarrow{\text{a.s.}} 0$$

where

$$\hat{S}_N \triangleq v(\gamma_N) \frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^* + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} v(\alpha_{i,n}) a_i a_i^*$$

with  $\gamma_N$  and  $\alpha_{1,n}, \dots, \alpha_{\varepsilon_n n, n}$  unique positive solutions to

$$\gamma_N = \frac{1}{N} \text{tr} C_N \left( \frac{(1-\varepsilon)v(\gamma_N)}{1 + cv(\gamma_N)\gamma_N} C_N + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} v(\alpha_{i,n}) a_i a_i^* \right)^{-1}$$
$$\alpha_{i,n} = \frac{1}{N} a_i^* \left( \frac{(1-\varepsilon)v(\gamma_N)}{1 + cv(\gamma_N)\gamma_N} C_N + \frac{1}{n} \sum_{j \neq i}^{\varepsilon_n n} v(\alpha_{j,n}) a_j a_j^* \right)^{-1} a_i, \quad i = 1, \dots, \varepsilon_n n.$$

- For  $\varepsilon_n n = 1$ ,

$$\hat{S}_N = v \left( \frac{\phi^{-1}(1)}{1-c} \right) \frac{1}{n} \sum_{i=1}^{n-1} x_i x_i^* + \left( v \left( \frac{\phi^{-1}(1)}{1-c} \frac{1}{N} a_1^* C_N^{-1} a_1 \right) + o(1) \right) a_1 a_1^*$$

Outlier rejection relies on  $\frac{1}{N} a_1^* C_N^{-1} a_1 \leq 1$ .

# Outlier Data

- For  $\varepsilon_n n = 1$ ,

$$\hat{S}_N = v \left( \frac{\phi^{-1}(1)}{1-c} \right) \frac{1}{n} \sum_{i=1}^{n-1} x_i x_i^* + \left( v \left( \frac{\phi^{-1}(1)}{1-c} \right) \frac{1}{N} a_1^* C_N^{-1} a_1 \right) + o(1) \Big) a_1 a_1^*$$

Outlier rejection relies on  $\frac{1}{N} a_1^* C_N^{-1} a_1 \leq 1$ .

- For  $a_i \sim \mathcal{CN}(0, D_N)$ ,  $\varepsilon_n \rightarrow \varepsilon \geq 0$ ,

$$\begin{aligned} \hat{S}_N &= v(\gamma_n) \frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^* + v(\alpha_n) \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} a_i a_i^* \\ \gamma_n &= \frac{1}{N} \text{tr} C_N \left( \frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n} C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n} D_N \right)^{-1} \\ \alpha_n &= \frac{1}{N} \text{tr} D_N \left( \frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n} C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n} D_N \right)^{-1}. \end{aligned}$$

## Outlier Data

- For  $\varepsilon_n n = 1$ ,

$$\hat{S}_N = v \left( \frac{\phi^{-1}(1)}{1-c} \right) \frac{1}{n} \sum_{i=1}^{n-1} x_i x_i^* + \left( v \left( \frac{\phi^{-1}(1)}{1-c} \frac{1}{N} a_1^* C_N^{-1} a_1 \right) + o(1) \right) a_1 a_1^*$$

Outlier rejection relies on  $\frac{1}{N} a_1^* C_N^{-1} a_1 \leq 1$ .

- For  $a_i \sim \mathcal{CN}(0, D_N)$ ,  $\varepsilon_n \rightarrow \varepsilon \geq 0$ ,

$$\begin{aligned} \hat{S}_N &= v(\gamma_n) \frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^* + v(\alpha_n) \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} a_i a_i^* \\ \gamma_n &= \frac{1}{N} \text{tr} C_N \left( \frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n} C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n} D_N \right)^{-1} \\ \alpha_n &= \frac{1}{N} \text{tr} D_N \left( \frac{(1-\varepsilon)v(\gamma_n)}{1+cv(\gamma_n)\gamma_n} C_N + \frac{\varepsilon v(\alpha_n)}{1+cv(\alpha_n)\alpha_n} D_N \right)^{-1}. \end{aligned}$$

For  $\varepsilon_n \rightarrow 0$ ,

$$\hat{S}_N = v \left( \frac{\phi^{-1}(1)}{1-c} \right) \frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^* + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} v \left( \frac{\phi^{-1}(1)}{1-c} \frac{1}{N} \text{tr} D_N C_N^{-1} \right) a_i a_i^*$$

Outlier rejection relies on  $\frac{1}{N} \text{tr} D_N C_N^{-1} \leq 1$ .

# Outlier Data

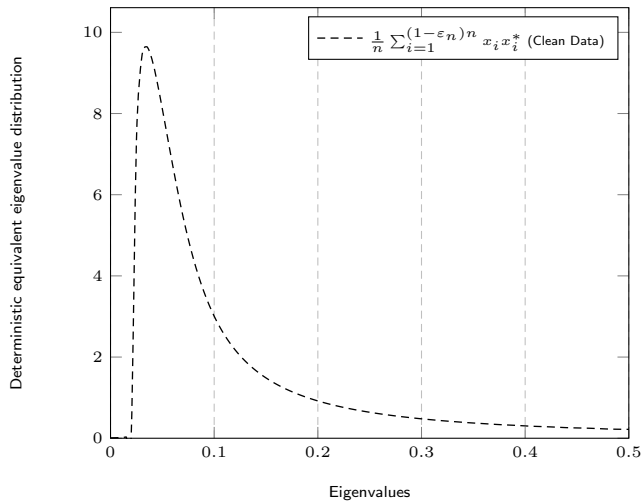


Figure: Limiting eigenvalue distributions.  $[C_N]_{ij} = .9^{|i-j|}$ ,  $D_N = I_N$ ,  $\varepsilon = .05$ .

# Outlier Data

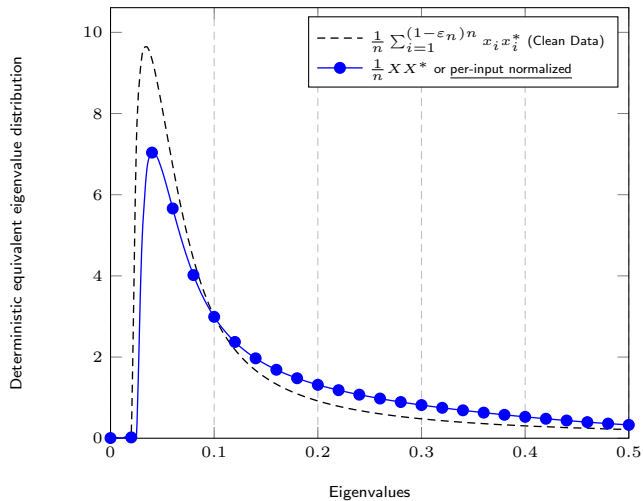


Figure: Limiting eigenvalue distributions.  $[C_N]_{ij} = .9^{|i-j|}$ ,  $D_N = I_N$ ,  $\varepsilon = .05$ .

# Outlier Data

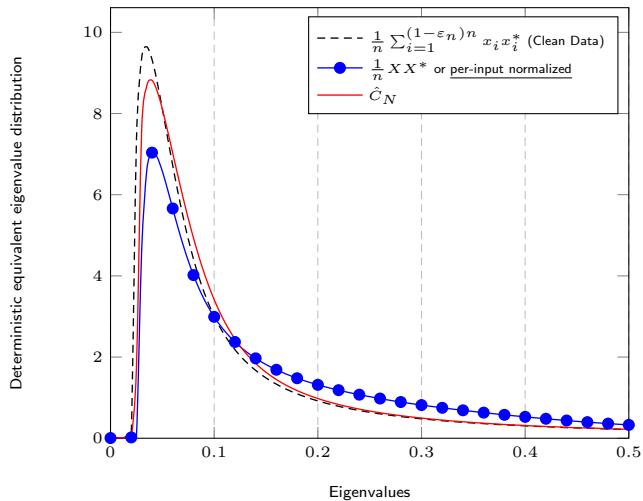


Figure: Limiting eigenvalue distributions.  $[C_N]_{ij} = .9^{|i-j|}$ ,  $D_N = I_N$ ,  $\varepsilon = .05$ .

Spectral Clustering Methods and Random Matrices

Community Detection on Graphs

Kernel Spectral Clustering

Kernel Spectral Clustering: Subspace Clustering

Semi-supervised Learning

Support Vector Machines

Neural Networks: Extreme Learning Machines

Random Matrices and Robust Estimation

**Perspectives**



# 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
- 💡 Study of 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

## 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.

## Summary of Results and Perspectives II



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.

## Summary of Results and Perspectives III



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

## Kernel methods.

- ✓ Subspace spectral clustering
- ✓ Subspace spectral clustering for  $f'(\tau) = 0$
- ✎ Spectral clustering with outer product kernel  $f(x^T y)$
- ✓ Semi-supervised learning, kernel approaches.
- ✓ Least square support vector machines (LS-SVM).
- ✎ Support vector machines (SVM).

## Applications.

- ✓ Massive MIMO user clustering

## References.



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



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.

## Summary of Results and Perspectives II



Z. Liao, R. Couillet, "Random matrices meet machine learning: a large dimensional analysis of LS-SVM", (submitted to) IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017.



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

## Community detection.

- ✓ Complete study of eigenvector contents in adjacency/modularity methods.
- 💡 Study of Bethe Hessian approach for the DCSBM model.
- 💡 Analysis of non-necessarily spectral approaches (wavelet approaches).

## 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

## Neural Networks.

- ✓ Non-linear extreme learning machines (ELM)
- ✎ Multi-layer ELM
- 💡 Backpropagation in ELM
- ✎ Random convolutional networks for image processing
- ✓ Linear echo-state networks (ESN)
- 💡 Non-linear ESN
- 💡 Connecting kernel methods to neural networks

## References.

-  C. Williams, "Computation with infinite neural networks", Neural Computation, 10(5), 1203-1216, 1998.
-  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, R. Couillet, "Harnessing neural networks: a random matrix approach", (submitted to) IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 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.

## Sparse PCA

- ✓ Spike random matrix sparse PCA
- ✎ Sparse kernel PCA

## References.



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

## Signal processing on graphs, distributed optimization, etc.

- 💡 Turning signal processing on graph methods random.
- 💡 Random matrix analysis of diffusion networks performance.



Thank you.