A Random Matrix Approach to Bigdata Machine Learning

(Statistics/Learning at Paris-Saclay (3rd Edition))

Romain COUILLET

LSS, CentraleSupélec
GSTATS DataScience Chair, University of Grenoble-Alps.

January, 2018

Outline

Outline

Random Matrices in Statistics

Motivation: $y_1, \ldots, y_n \in \mathbb{R}^p$ i.i.d. with $y_i \sim \mathcal{N}(0, C_p)$:

Random Matrices in Statistics

Motivation: $y_1, \ldots, y_n \in \mathbb{R}^p$ i.i.d. with $y_i \sim \mathcal{N}(0, C_p)$:

- Maximum likelihood estimator for C_p:

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

with $Y_p = [y_1, \ldots, y_n] \in \mathbb{R}^{p \times n}$

Random Matrices in Statistics

Motivation: $y_1, \ldots, y_n \in \mathbb{R}^p$ i.i.d. with $y_i \sim \mathcal{N}(0, C_p)$:

- Maximum likelihood estimator for C_p:

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^\mathsf{T} = \frac{1}{n} Y_p Y_p^\mathsf{T} = \frac{1}{n} C_p^{\frac{1}{2}} X_p X_p^\mathsf{T} C_p^{\frac{1}{2}}$$

with $Y_p = [y_1, \ldots, y_n] \in \mathbb{R}^{p \times n}$ and $[X_p]_{ij} \sim \mathcal{N}(0, 1)$ i.i.d.

Random Matrices in Statistics

Motivation: $y_1, \ldots, y_n \in \mathbb{R}^p$ i.i.d. with $y_i \sim \mathcal{N}(0, C_p)$:

- Maximum likelihood estimator for C_p:

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^\mathsf{T} = \frac{1}{n} Y_p Y_p^\mathsf{T} = \frac{1}{n} C_p^{\frac{1}{2}} X_p X_p^\mathsf{T} C_p^{\frac{1}{2}}$$

  with $Y_p = [y_1, \ldots, y_n] \in \mathbb{R}^{p \times n}$ and $[X_p]_{ij} \sim \mathcal{N}(0,1)$ i.i.d.

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

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

Random Matrices in Statistics

Motivation: $y_1, \ldots, y_n \in \mathbb{R}^p$ i.i.d. with $y_i \sim \mathcal{N}(0, C_p)$:

- Maximum likelihood estimator for C_p:

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^\mathsf{T} = \frac{1}{n} Y_p Y_p^\mathsf{T} = \frac{1}{n} C_p^{\frac{1}{2}} X_p X_p^\mathsf{T} C_p^{\frac{1}{2}}$$

  with $Y_p = [y_1, \ldots, y_n] \in \mathbb{R}^{p \times n}$ and $[X_p]_{ij} \sim \mathcal{N}(0,1)$ i.i.d.
- If $n \to \infty$, then, by strong law of large numbers

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

Random Matrix Regime

- Becomes wrong when $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

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

Random Matrices in Statistics

Motivation: $y_1, \dots, y_n \in \mathbb{R}^p$ i.i.d. with $y_i \sim \mathcal{N}(0, C_p)$:

- Maximum likelihood estimator for C_p:

$$\hat{C}_p = \frac{1}{n} \sum_{i=1}^{n} y_i y_i^\mathsf{T} = \frac{1}{n} Y_p Y_p^\mathsf{T} = \frac{1}{n} C_p^{\frac{1}{2}} X_p X_p^\mathsf{T} C_p^{\frac{1}{2}}$$

  with $Y_p = [y_1, \dots, y_n] \in \mathbb{R}^{p \times n}$ and $[X_p]_{ij} \sim \mathcal{N}(0,1)$ i.i.d.

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

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

Random Matrix Regime

- Becomes wrong when $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

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

If $p \simeq n$, **severe consequences** for estimators/algorithms based on $\hat{C}_p$.

Global behavior: the Marčenko–Pastur law (1967)

- Assume first $C_p = I_p$.

Global behavior: the Marčenko–Pastur law (1967)

- Assume first $C_p = I_p$.

For $n, p \to \infty$ with $p/n \to c$, "weak" convergence of eigenvalue distribution

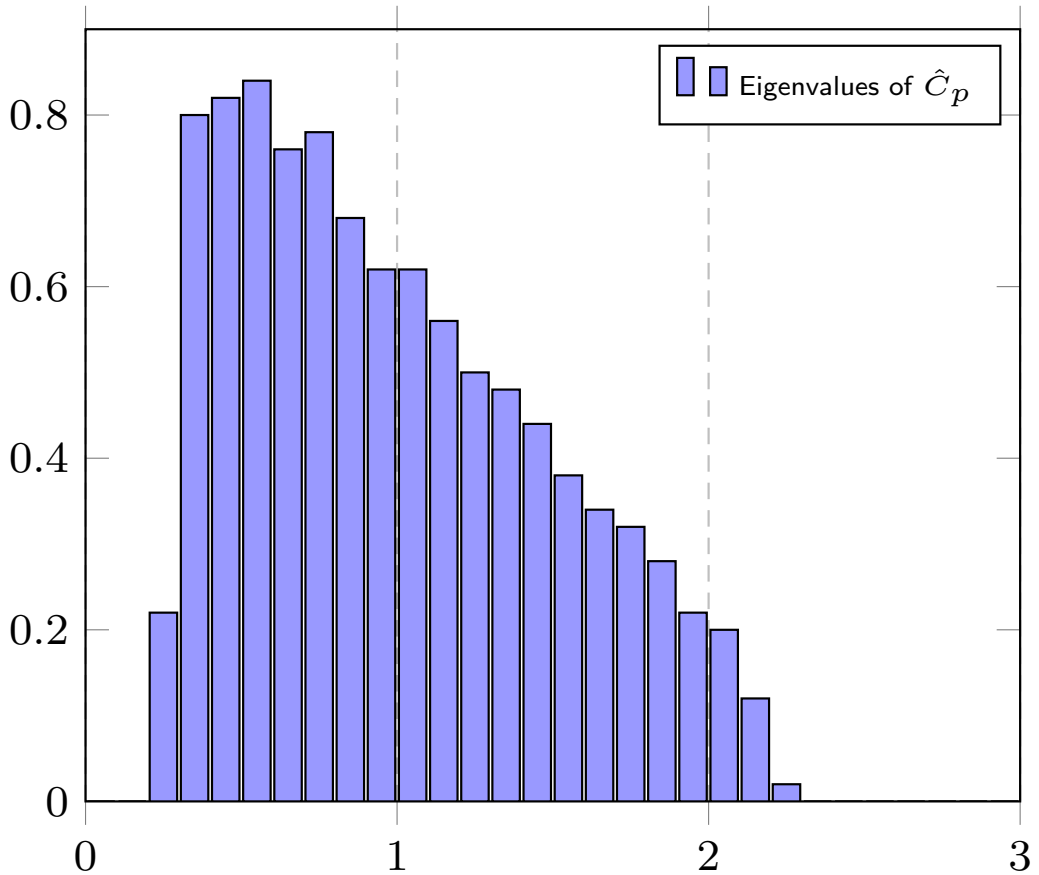


Figure: Histogram of eigenvalues of $\hat{C}_p$ for $p = 500$, $n = 2000$, $C_p = I_p$.

Global behavior: the Marčenko–Pastur law (1967)

- Assume first $C_p = I_p$.

For $n, p \to \infty$ with $p/n \to c$, "weak" convergence of eigenvalue distribution

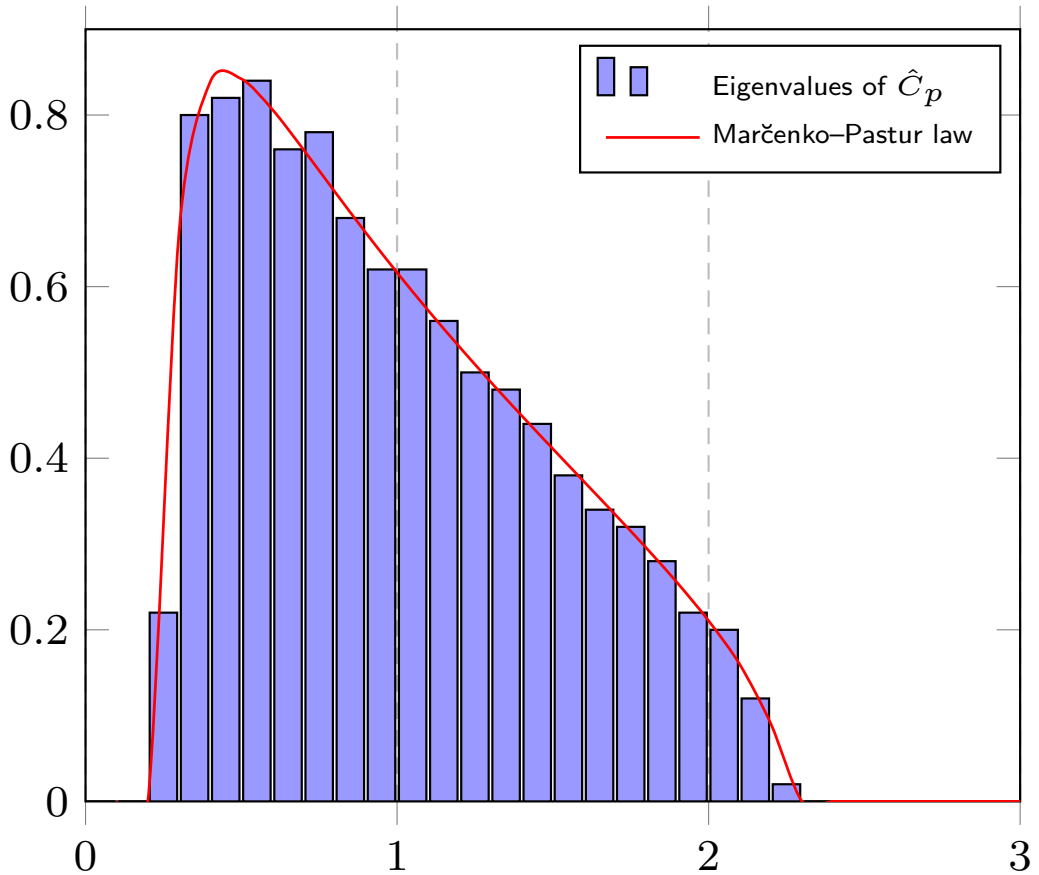


Figure: Histogram of eigenvalues of $\hat{C}_p$ for $p = 500$, $n = 2000$, $C_p = I_p$.

Local behavior: Bai–Silverstein theorem (1998)

(additional assumption: $E[|[X_p]_{ij}|^4] < \infty$*)*

Local behavior: Bai–Silverstein theorem (1998)

(additional assumption: $E[|[X_p]_{ij}|^4] < \infty$*)*

For $C_p = I_p$, eigenvalues remain concentrated

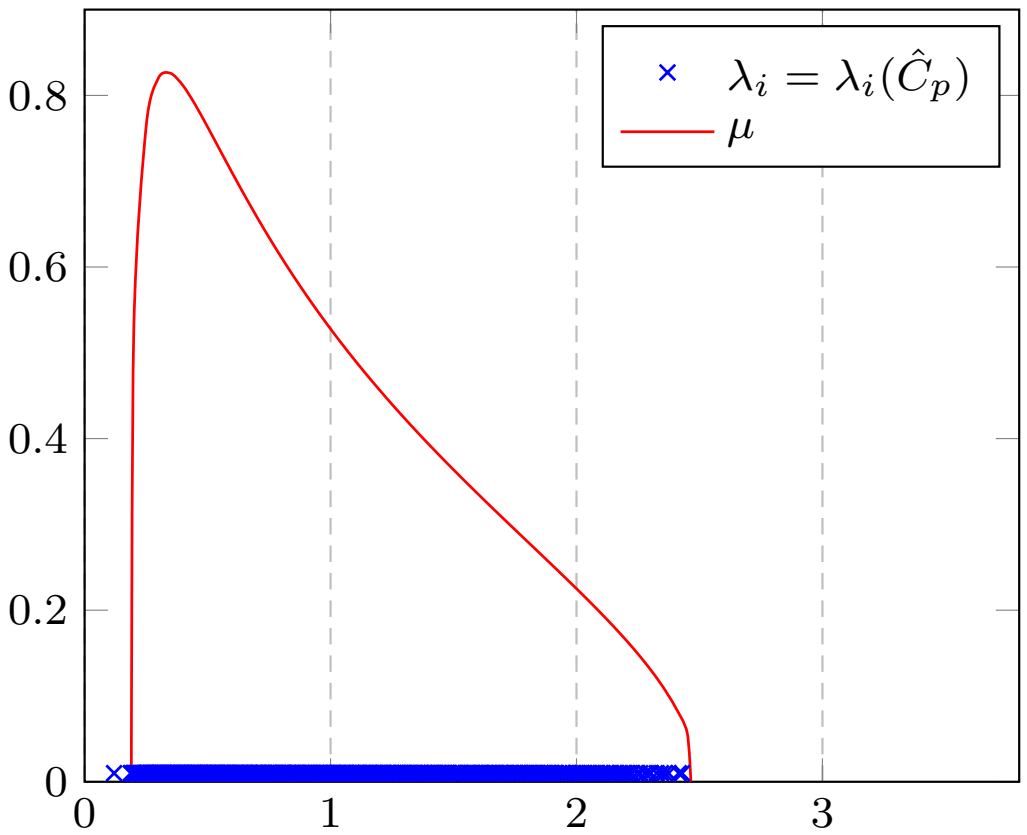


Figure: Eigenvalues of $\hat{C}_p$ for $C_p = I_p$, $p = 500$, $n = 1500$.

Spiked models

Assume $C_p = I_p + \sum_{i=1}^{k} \omega_i u_i u_i^\mathsf{T}$ with k **small**, and $\hat{C}_p = \sum_{i=1}^{p} \lambda_i \hat{u}_i \hat{u}_i^\mathsf{T}$

Spiked models

Assume $C_p = I_p + \sum_{i=1}^k \omega_i u_i u_i^\mathsf{T}$ with k **small**, and $\hat{C}_p = \sum_{i=1}^p \lambda_i \hat{u}_i \hat{u}_i^\mathsf{T}$

Phase transition **iif** $\omega_i > \sqrt{c}$

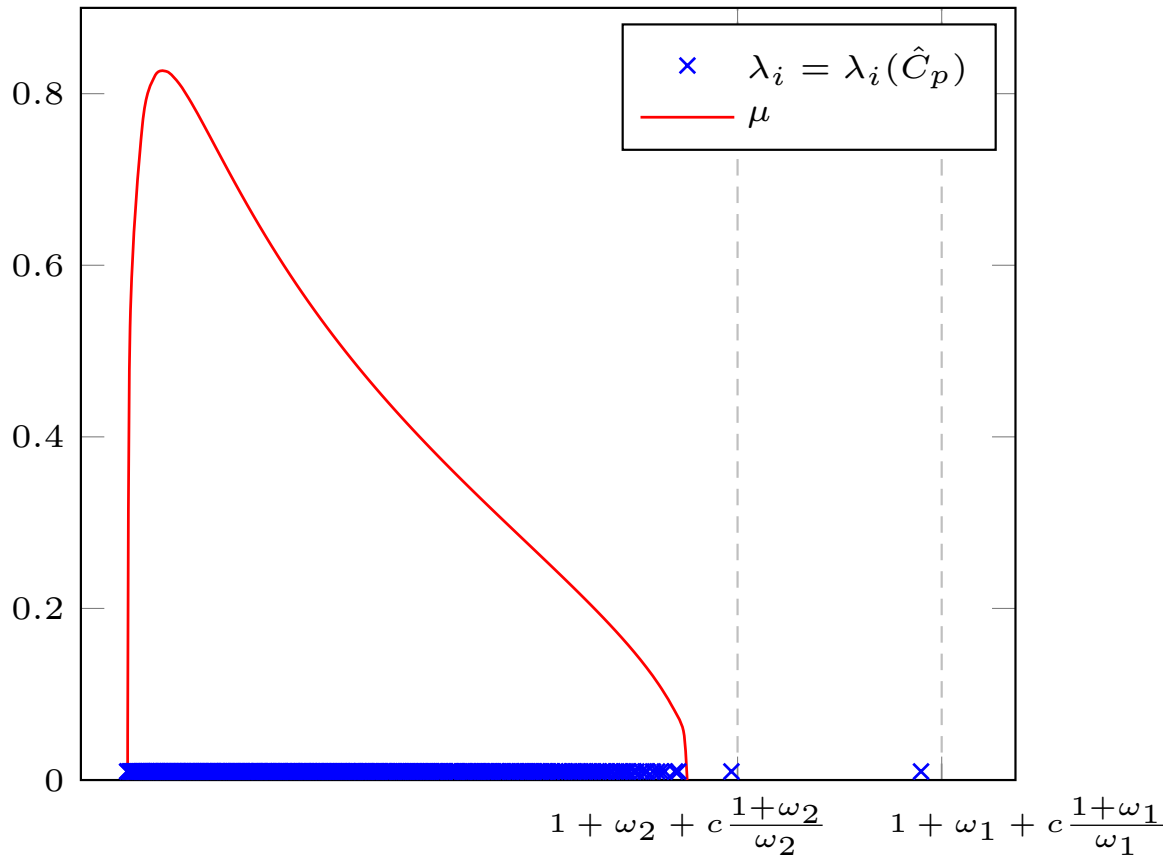


Figure: Eigenvalues of $\frac{1}{n} Y_p Y_p^*$, $C_p = I_p + \underbrace{2}_{\omega_1} e_1 e_1^\mathsf{T} + \underbrace{1}_{\omega_2} e_2 e_2^\mathsf{T} + \underbrace{\frac{1}{2}}_{\omega_3} e_3 e_3^\mathsf{T}$, $p/n = \frac{1}{3}$.

Spiked models and eigenvectors

Assume $C_p = I_p + \sum_{i=1}^{k} \omega_i u_i u_i^\mathsf{T}$ with k **small**, and $\hat{C}_p = \sum_{i=1}^{p} \lambda_i \hat{u}_i \hat{u}_i^\mathsf{T}$

Spiked models and eigenvectors

Assume $C_p = I_p + \sum_{i=1}^{k} \omega_i u_i u_i^\mathsf{T}$ with k **small**, and $\hat{C}_p = \sum_{i=1}^{p} \lambda_i \hat{u}_i \hat{u}_i^\mathsf{T}$

For $C_p = I_p + P$, phase transition $\omega_i > \sqrt{c}$ on alignment of u_i and $\hat{u}_i$

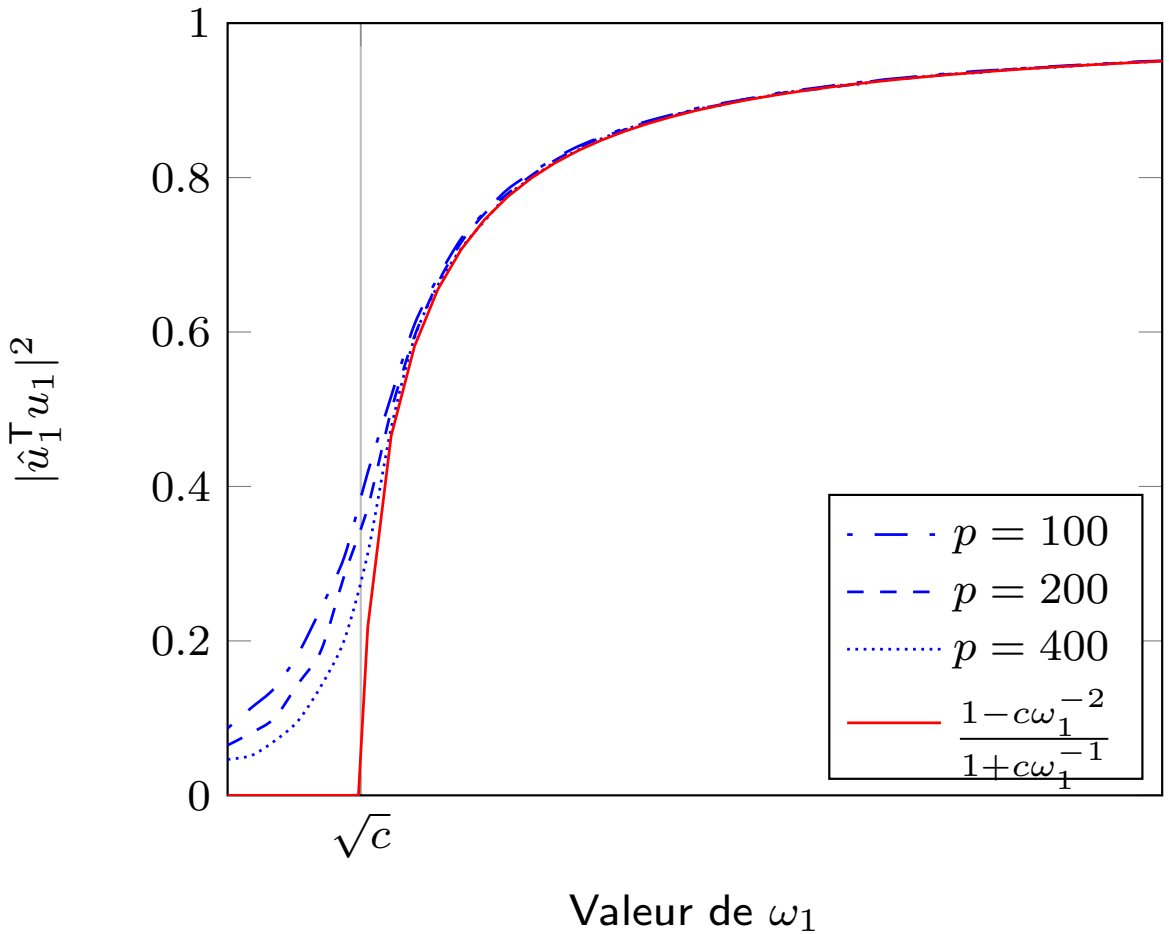


Figure: Alignment $|\hat{u}_1^\mathsf{T} u_1|^2$ for $C_p = I_p + \omega_1 u_1 u_1^\mathsf{T}$, $p/n \to c = 1/3$.

Spiked models

Available for many models (with P of **small rank**):

- $\hat{C}_p = \frac{1}{n}(I_p + P)^{\frac{1}{2}} X_p X_p^{\mathsf{T}} (I_p + P)^{\frac{1}{2}}$

Spiked models

Available for many models (with P of **small rank**):

- $\hat{C}_p = \frac{1}{n}(I_p + P)^{\frac{1}{2}} X_p X_p^\mathsf{T} (I_p + P)^{\frac{1}{2}}$
- $\hat{C}_p = \frac{1}{n} X_p X_p^\mathsf{T} + P$
- $\hat{C}_p = \frac{1}{n} X_p^\mathsf{T} (I + P) X_p$
- $\hat{C}_p = \frac{1}{n}(X_p + P)^\mathsf{T}(X_p + P)$
- etc.

Spiked models

Available for many models (with P of **small rank**):

- $\hat{C}_p = \frac{1}{n}(I_p + P)^{\frac{1}{2}} X_p X_p^\mathsf{T} (I_p + P)^{\frac{1}{2}}$
- $\hat{C}_p = \frac{1}{n} X_p X_p^\mathsf{T} + P$
- $\hat{C}_p = \frac{1}{n} X_p^\mathsf{T} (I + P) X_p$
- $\hat{C}_p = \frac{1}{n}(X_p + P)^\mathsf{T}(X_p + P)$
- etc.

Applications:

- Evaluation of algorithm performances in large dimensions (detection, estimation, subspace methods, etc.)

Spiked models

Available for many models (with P of **small rank**):

- $\hat{C}_p = \frac{1}{n}(I_p + P)^{\frac{1}{2}} X_p X_p^\mathsf{T} (I_p + P)^{\frac{1}{2}}$
- $\hat{C}_p = \frac{1}{n} X_p X_p^\mathsf{T} + P$
- $\hat{C}_p = \frac{1}{n} X_p^\mathsf{T} (I + P) X_p$
- $\hat{C}_p = \frac{1}{n}(X_p + P)^\mathsf{T}(X_p + P)$
- etc.

Applications:

- Evaluation of algorithm performances in large dimensions (detection, estimation, subspace methods, etc.)
- Improved algorithms

Spiked models

Available for many models (with P of **small rank**):

- $\hat{C}_p = \frac{1}{n}(I_p + P)^{\frac{1}{2}} X_p X_p^\mathsf{T} (I_p + P)^{\frac{1}{2}}$
- $\hat{C}_p = \frac{1}{n} X_p X_p^\mathsf{T} + P$
- $\hat{C}_p = \frac{1}{n} X_p^\mathsf{T} (I + P) X_p$
- $\hat{C}_p = \frac{1}{n}(X_p + P)^\mathsf{T}(X_p + P)$
- etc.

Applications:

- Evaluation of algorithm performances in large dimensions (detection, estimation, subspace methods, etc.)
- Improved algorithms
- Applications in statistics, biostats, finance, signal processing, etc.

Outline

Models in Machine Learning

"Standard" results valid for random matrix models with:

- independent or linearly dependent entries
- "Gaussian-like" data.

Models in Machine Learning

"Standard" results valid for random matrix models with:

- independent or linearly dependent entries
- "Gaussian-like" data.

In machine learning:

- kernel methods, neural networks, all non linear
- models with outliers, structured data, multi-modal, heterogeneous, etc.

Models in Machine Learning

"Standard" results valid for random matrix models with:

- independent or linearly dependent entries
- "Gaussian-like" data.

In machine learning:

- kernel methods, neural networks, all non linear
- models with outliers, structured data, multi-modal, heterogeneous, etc.

Question. What happens to classical machine learning methods when $n, p \to \infty$?

Models in Machine Learning

"Standard" results valid for random matrix models with:

- independent or linearly dependent entries
- "Gaussian-like" data.

In machine learning:

- kernel methods, neural networks, all non linear
- models with outliers, structured data, multi-modal, heterogeneous, etc.

Question. What happens to classical machine learning methods when $n, p \to \infty$?

Objective: Understand and improve statistical learning methods in large dimensions.

Models in Machine Learning

"Standard" results valid for random matrix models with:

- independent or linearly dependent entries
- "Gaussian-like" data.

In machine learning:

- kernel methods, neural networks, all non linear
- models with outliers, structured data, multi-modal, heterogeneous, etc.

Question. What happens to classical machine learning methods when $n, p \to \infty$?

Objective: Understand and improve statistical learning methods in large dimensions.

$\longrightarrow$ so to better understand the core algorithms (huge need for industries)

Models in Machine Learning

"Standard" results valid for random matrix models with:

- independent or linearly dependent entries
- "Gaussian-like" data.

In machine learning:

- kernel methods, neural networks, all non linear
- models with outliers, structured data, multi-modal, heterogeneous, etc.

Question. What happens to classical machine learning methods when $n, p \to \infty$?

Objective: Understand and improve statistical learning methods in large dimensions.

$\longrightarrow$ so to better understand the core algorithms (huge need for industries)

$\longrightarrow$ so to improve advanced **practical** methods

Outline

Kernel Methods: the results

Context: Clustering (i.e., unsupervised classification) of large data $x_1, \ldots, x_n \in \mathbb{R}^p$.

Kernel Methods: the results

Context: Clustering (i.e., unsupervised classification) of large data $x_1, \ldots, x_n \in \mathbb{R}^p$.

Kernel spectral classification methods: Information in dominant eigenvectors of

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

Kernel Methods: the results

Context: Clustering (i.e., unsupervised classification) of large data $x_1, \ldots, x_n \in \mathbb{R}^p$.

Kernel spectral classification methods: Information in dominant eigenvectors of

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

The results:

- k-class non-trivial Gaussian mixture for x_i ($\|\mu_a - \mu_b\| = O_p(1)$, $\|C_a - C_b\| = O_p(1)$); class sizes $n_1 + \ldots + n_k = n$.
- large p, n, n_i, small k.

Kernel Methods: the results

Context: Clustering (i.e., unsupervised classification) of large data $x_1, \ldots, x_n \in \mathbb{R}^p$.

Kernel spectral classification methods: Information in dominant eigenvectors of

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

The results:

- k-class non-trivial Gaussian mixture for x_i ($\|\mu_a - \mu_b\| = O_p(1)$, $\|C_a - C_b\| = O_p(1)$); class sizes $n_1 + \ldots + n_k = n$.
- large p, n, n_i, small k.
- It can be shown that

$$\left\| K - \tilde{K} \right\| \xrightarrow{\text{a.s.}} 0, \text{ with } \tilde{K} \text{ **linear spiked model**.}$$

Kernel Methods: the results

Context: Clustering (i.e., unsupervised classification) of large data $x_1, \ldots, x_n \in \mathbb{R}^p$.

Kernel spectral classification methods: Information in dominant eigenvectors of

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

The results:

- k-class non-trivial Gaussian mixture for x_i ($\|\mu_a - \mu_b\| = O_p(1)$, $\|C_a - C_b\| = O_p(1)$); class sizes $n_1 + \ldots + n_k = n$.
- large p, n, n_i, small k.
- It can be shown that

$$\left\| K - \tilde{K} \right\| \xrightarrow{\text{a.s.}} 0, \text{ with } \tilde{K} \textbf{ linear spiked model}.$$

(key result: $\forall i, j,\ \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$, independently of classes).

Kernel Methods: the results

Context: Clustering (i.e., unsupervised classification) of large data $x_1, \ldots, x_n \in \mathbb{R}^p$.

Kernel spectral classification methods: Information in dominant eigenvectors of

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

The results:

- k-class non-trivial Gaussian mixture for x_i ($\|\mu_a - \mu_b\| = O_p(1)$, $\|C_a - C_b\| = O_p(1)$); class sizes $n_1 + \ldots + n_k = n$.
- large p, n, n_i, small k.
- It can be shown that

$$\left\| K - \tilde{K} \right\| \xrightarrow{\text{a.s.}} 0, \text{ with } \tilde{K} \text{ **linear spiked model**.}$$

  (key result: $\forall i, j,\ \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$, independently of classes).
- Roughly speaking,

$$\tilde{K} = \frac{1}{n} Z Z^\mathsf{T} + J B J^\mathsf{T} + *$$

  with $J = [j_1, \ldots, j_k]$, $j_i^\mathsf{T} = (0, \ldots, 0, 1_{n_i}, 0, \ldots, 0)$
  and B function of $f(\tau), f'(\tau), f''(\tau)$ and the mixture model means/covariances.

Kernel Methods: the results

Context: Clustering (i.e., unsupervised classification) of large data $x_1, \ldots, x_n \in \mathbb{R}^p$.

Kernel spectral classification methods: Information in dominant eigenvectors of

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

The results:

- k-class non-trivial Gaussian mixture for x_i ($\|\mu_a - \mu_b\| = O_p(1)$, $\|C_a - C_b\| = O_p(1)$); class sizes $n_1 + \ldots + n_k = n$.
- large p, n, n_i, small k.
- It can be shown that

$$\left\| K - \tilde{K} \right\| \xrightarrow{\text{a.s.}} 0, \text{ with } \tilde{K} \textbf{ linear spiked model}.$$

  (key result: $\forall i, j,\ \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau$, independently of classes).
- Roughly speaking,

$$\tilde{K} = \frac{1}{n} Z Z^\mathsf{T} + J B J^\mathsf{T} + *$$

  with $J = [j_1, \ldots, j_k]$, $j_i^\mathsf{T} = (0, \ldots, 0, 1_{n_i}, 0, \ldots, 0)$
  and B function of $f(\tau), f'(\tau), f''(\tau)$ and the mixture model means/covariances.
- **Surprising coincidence with real data.**

Theory versus MNIST dataset

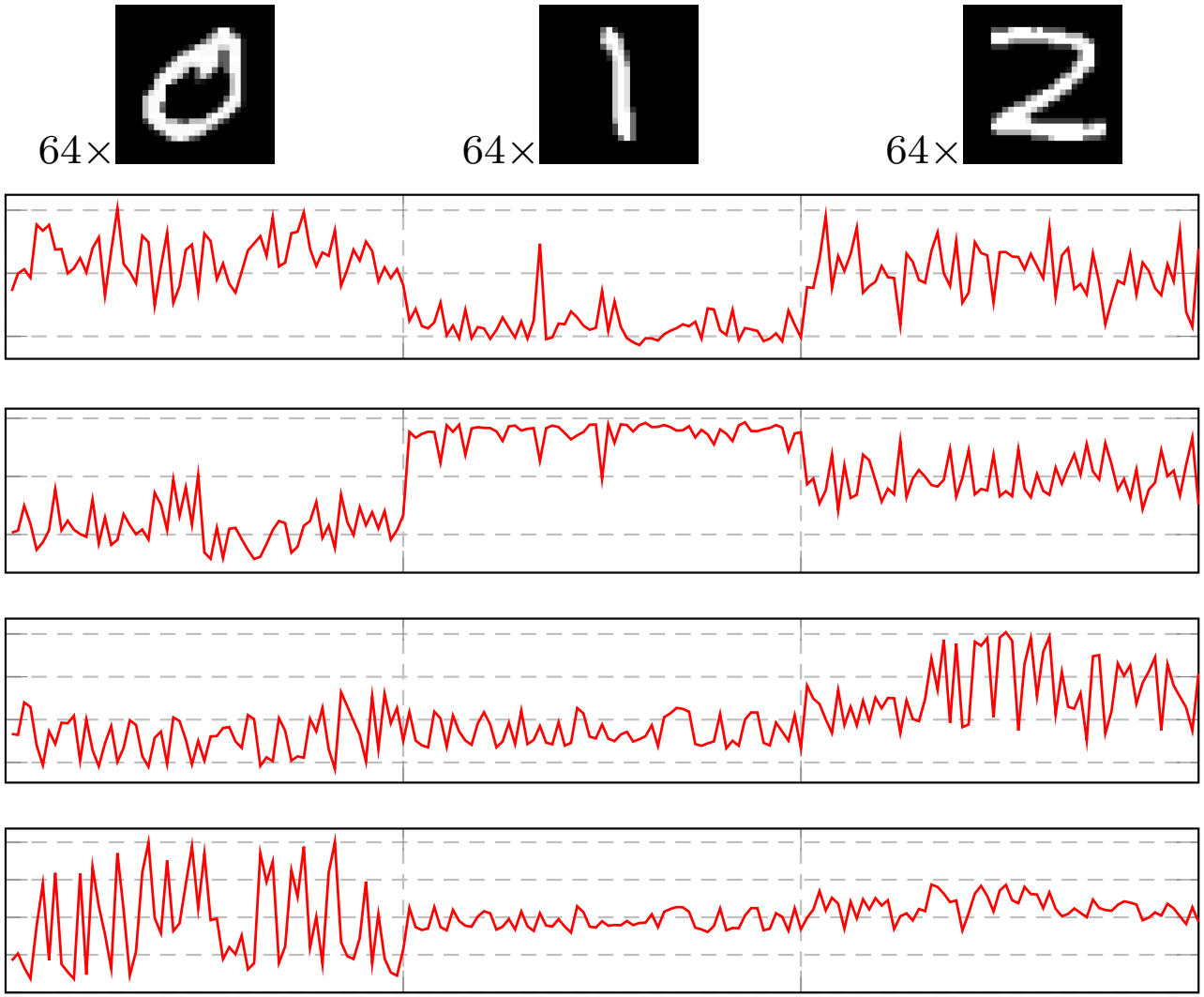


Figure: Dominant eigenvector of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$, MNIST (**red**) and theory based on $\tilde{K}$ (**blue**), kernel $f(\|x-y\|^2) = \exp(-\frac{1}{2}\|x-y\|^2)$, $n = 192 = 3 \times 64$, $p = 784$ (28×28).

Theory versus MNIST dataset

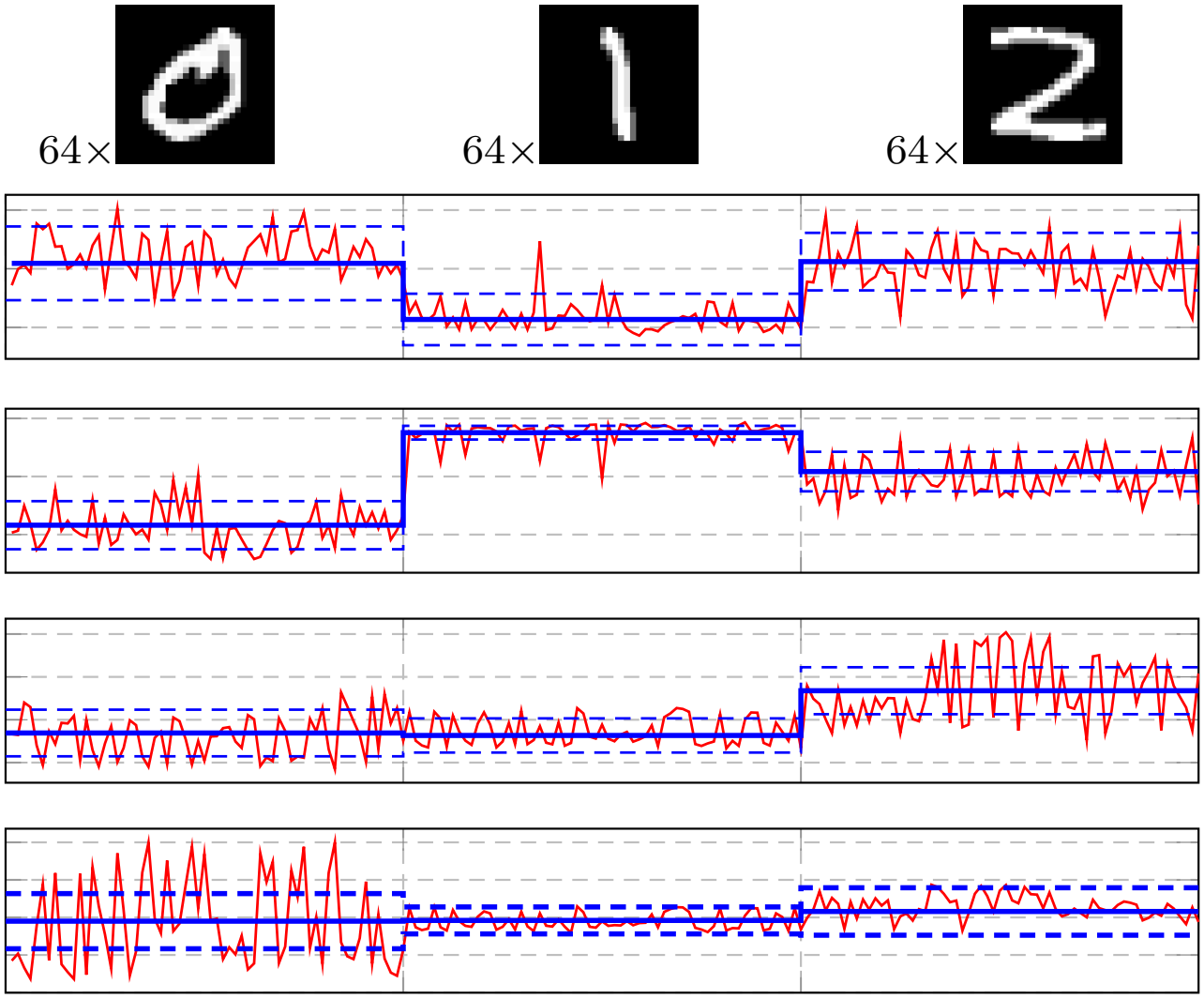
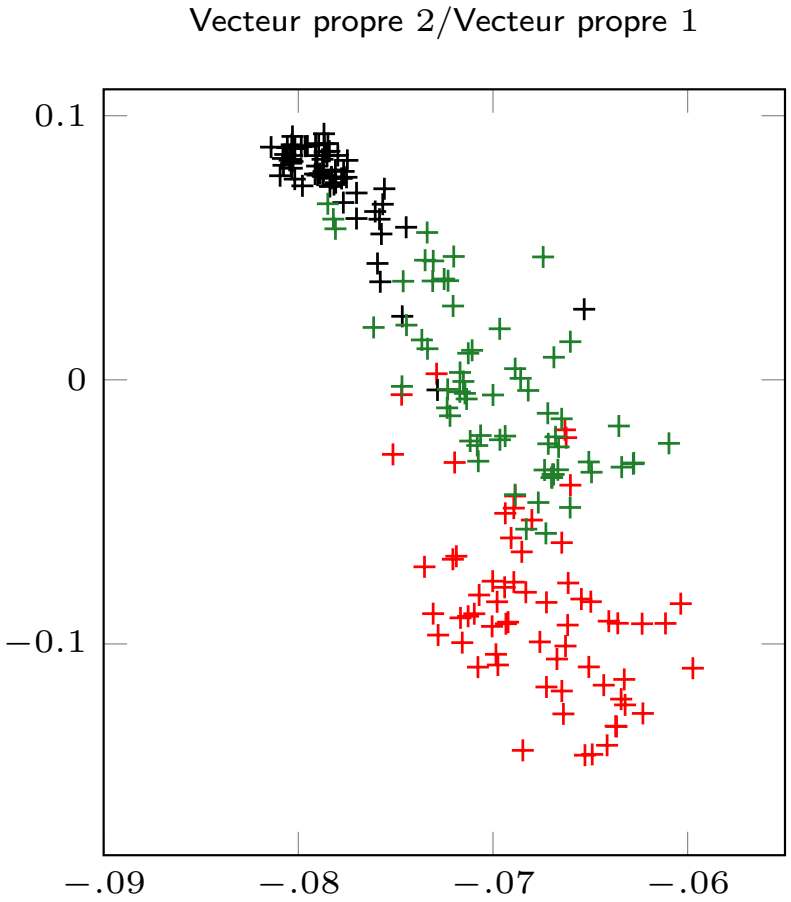
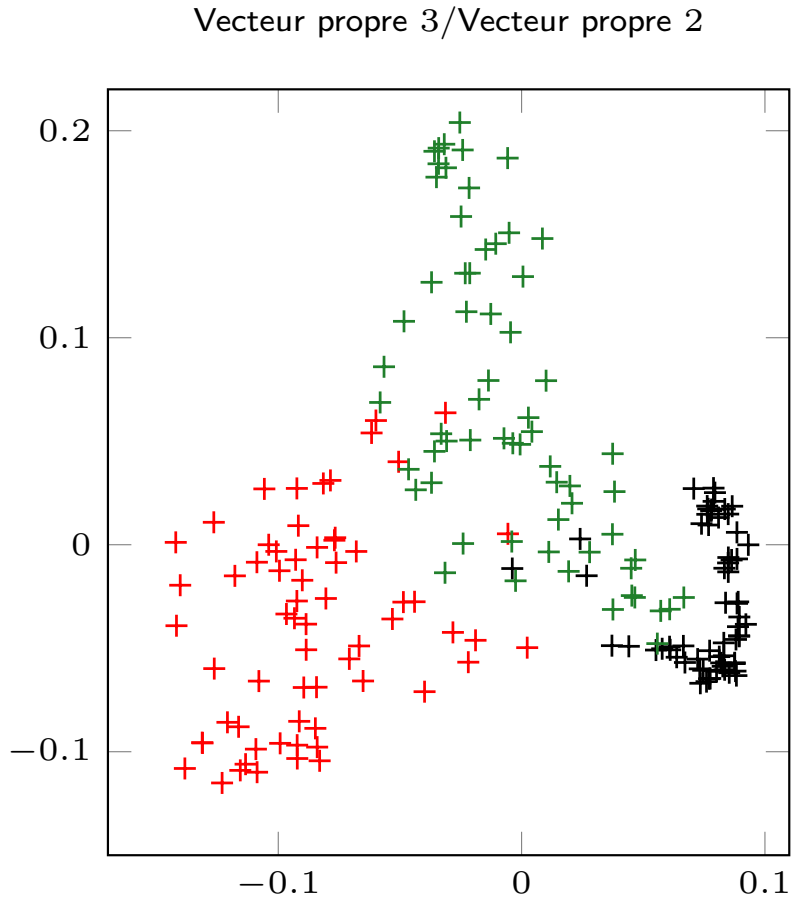


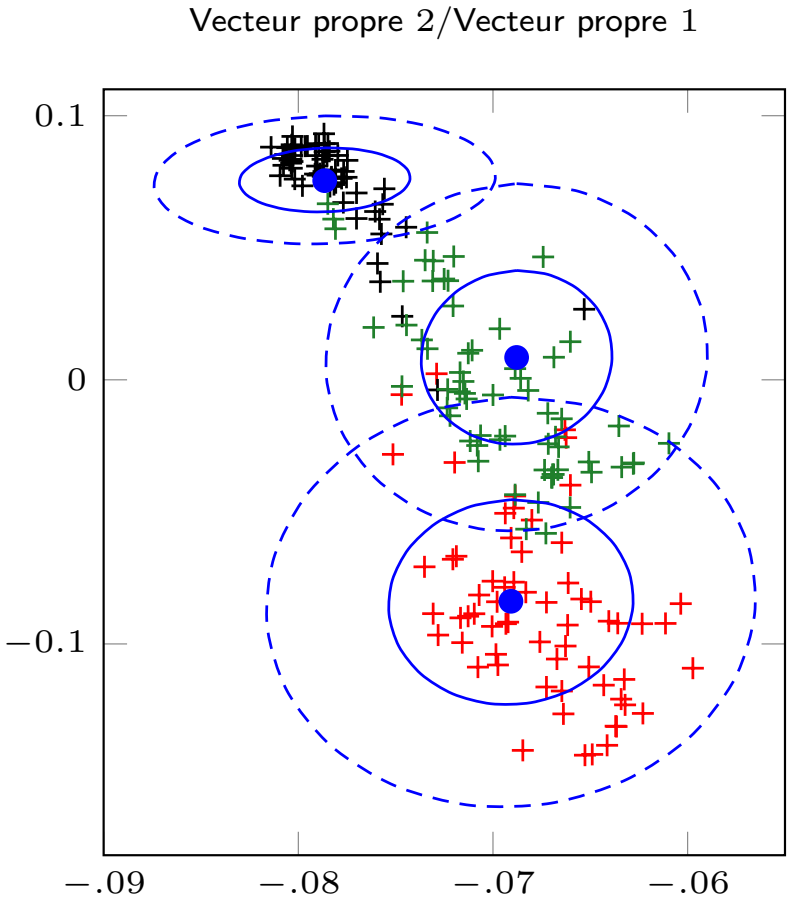
Figure: Dominant eigenvector of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$, MNIST (**red**) and theory based on $\tilde{K}$ (**blue**), kernel $f(\|x-y\|^2) = \exp(-\frac{1}{2}\|x-y\|^2)$, $n = 192 = 3 \times 64$, $p = 784$ (28×28).

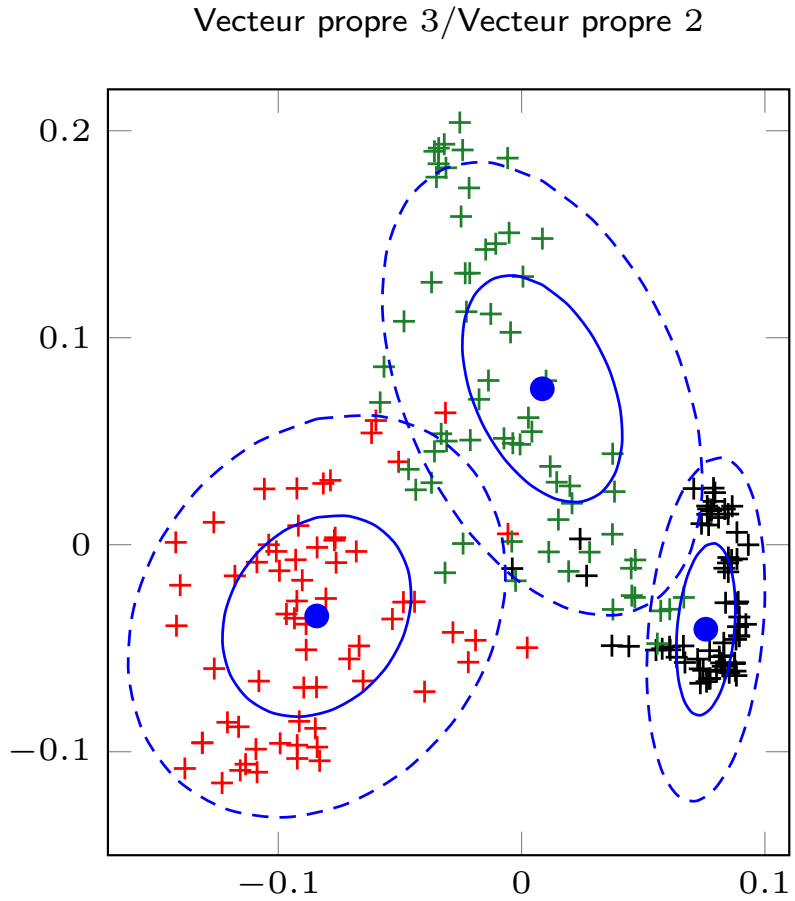
Theory and MNIST dataset





Theory and MNIST dataset





Practical consequences: the surprising $f'(\tau) = 0$ scenario

For data with close means, $f'(\tau) = 0$ is optimal

($\|C_a - C_b\| = O(p^{-\frac{1}{2}})$ achievable!)

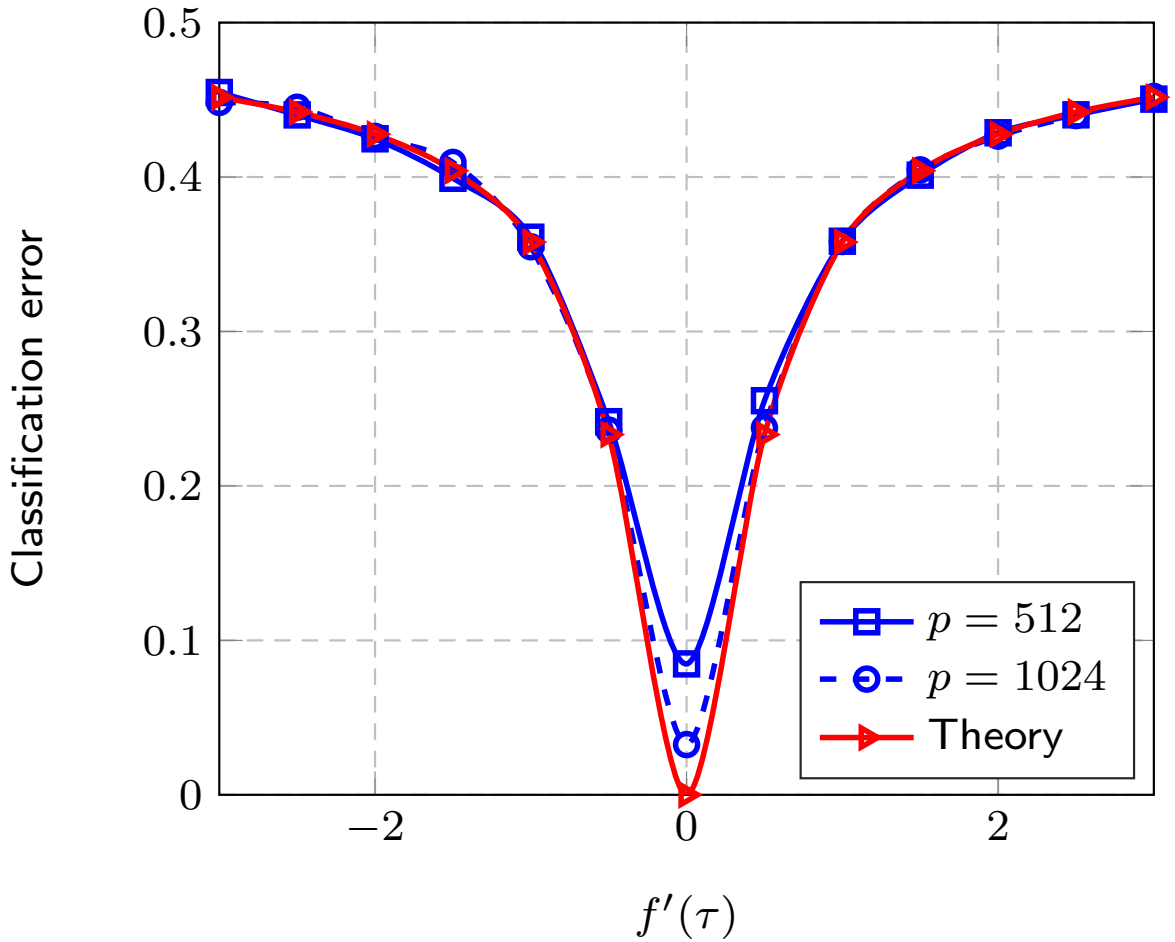


Figure: $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \sim \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

Outline

Community detection: the "DC-SBM" model

Stochastic Block Model (SBM): Adjacency $A_{ij} \sim \text{Bernoulli}(C_{g_i g_j})$
($g_i =$"community of i").
Degree Corrected-SBM (DC-SBM): Adjacency $A_{ij} \sim \text{Bernoulli}(q_i q_j C_{g_i g_j})$.

Community detection: the "DC-SBM" model

Stochastic Block Model (SBM): Adjacency $A_{ij} \sim \text{Bernoulli}(C_{g_i g_j})$
($g_i =$"community of i").
Degree Corrected-SBM (DC-SBM): Adjacency $A_{ij} \sim \text{Bernoulli}(q_i q_j C_{g_i g_j})$.

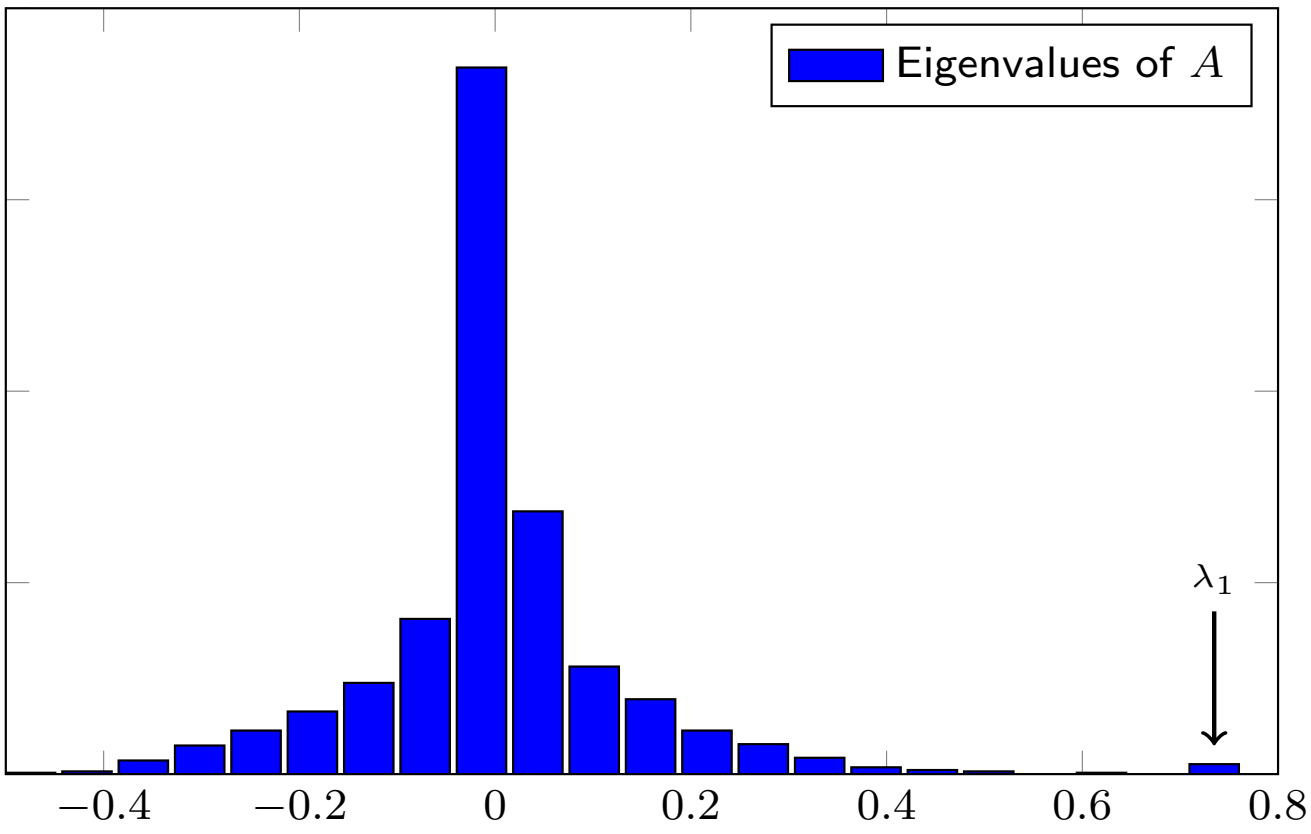


Figure: PolBlogs graph ($n \sim 1200$, two classes).

Community detection: the "DC-SBM" model

Stochastic Block Model (SBM): Adjacency $A_{ij} \sim \text{Bernoulli}(C_{g_i g_j})$
($g_i =$"community of i").
Degree Corrected-SBM (DC-SBM): Adjacency $A_{ij} \sim \text{Bernoulli}(q_i q_j C_{g_i g_j})$.

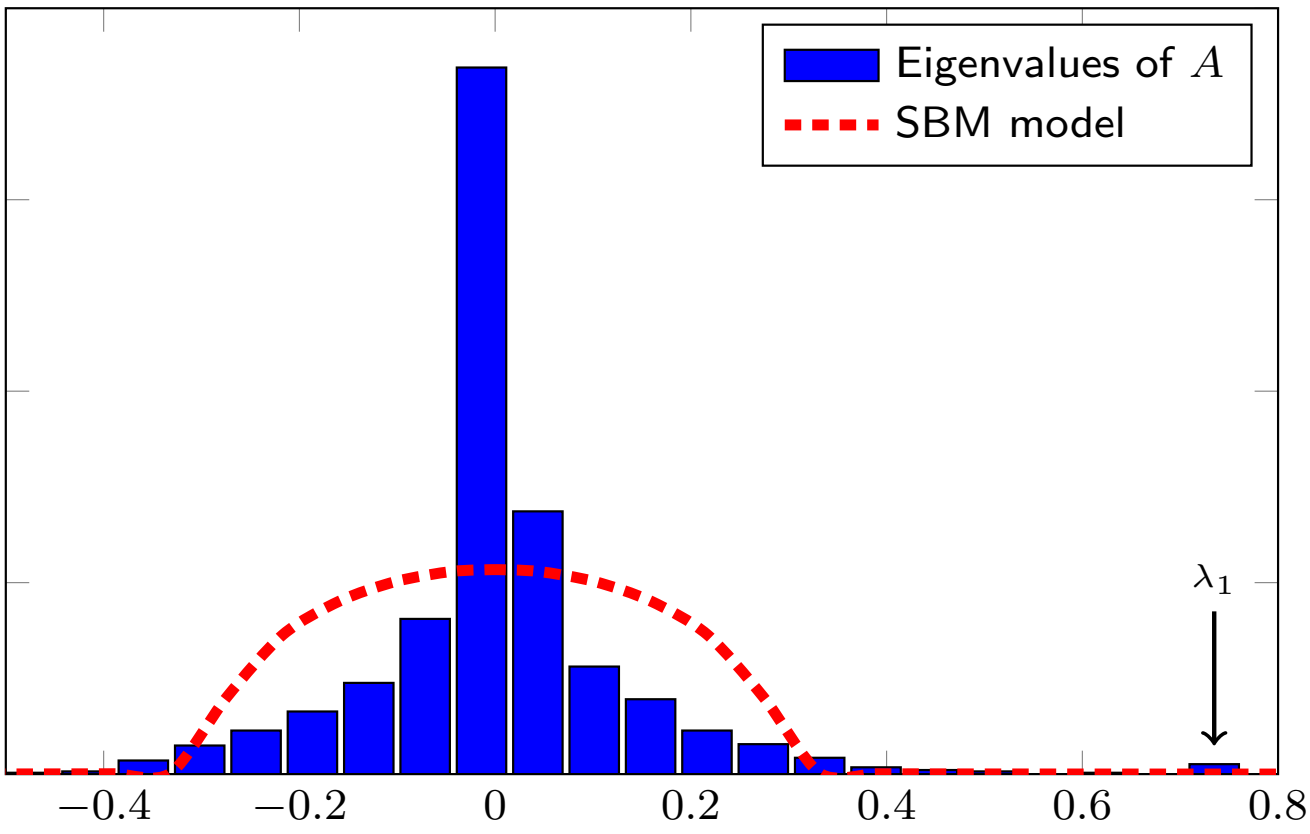


Figure: PolBlogs graph ($n \sim 1200$, two classes).

Community detection: the "DC-SBM" model

Stochastic Block Model (SBM): Adjacency $A_{ij} \sim \text{Bernoulli}(C_{g_i g_j})$
($g_i =$"community of i").
Degree Corrected-SBM (DC-SBM): Adjacency $A_{ij} \sim \text{Bernoulli}(q_i q_j C_{g_i g_j})$.

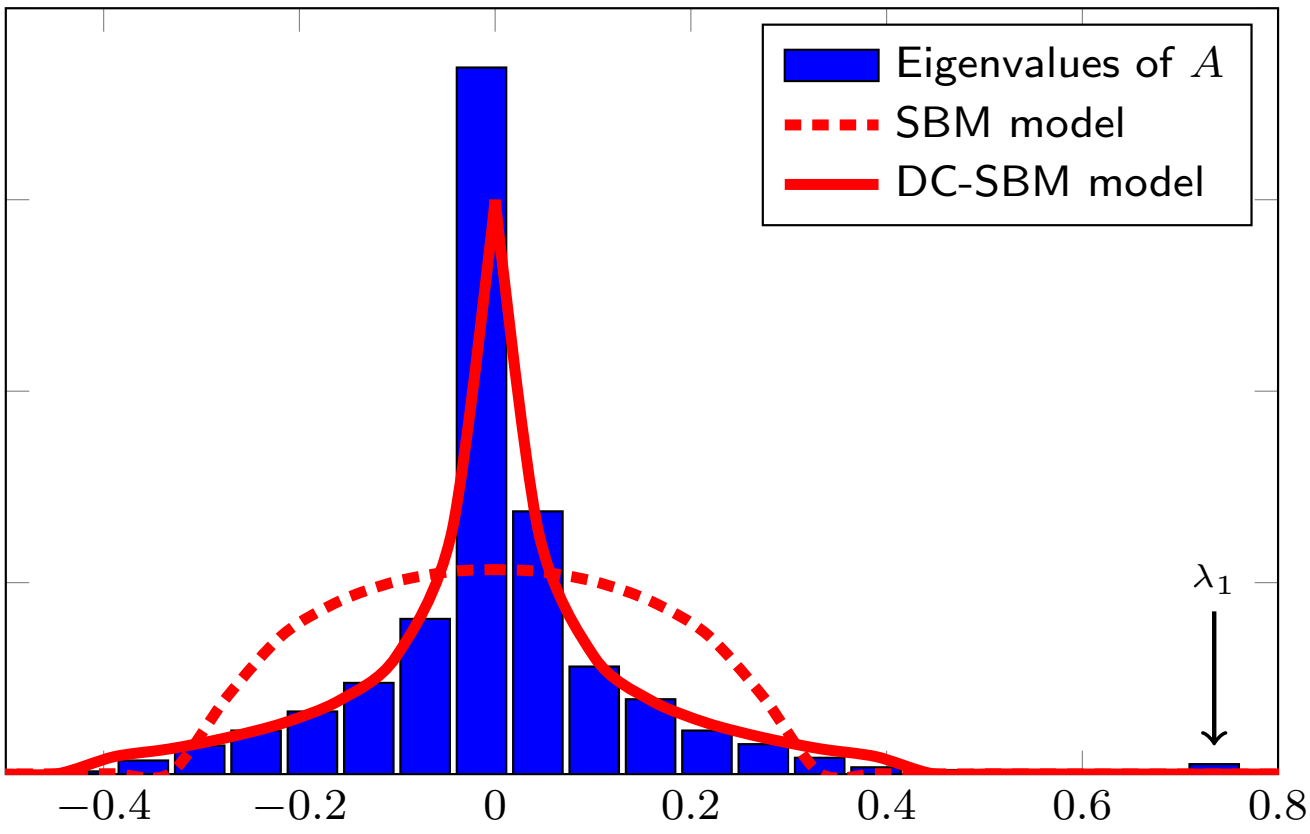


Figure: PolBlogs graph ($n \sim 1200$, two classes).

Community Detection: improved approach

Generalized Laplacian: $L_\alpha = D^{-\alpha} A D^{-\alpha}$

Community Detection: improved approach

Generalized Laplacian: $L_\alpha = D^{-\alpha} A D^{-\alpha}$

Theoretical Results:

- Support of the eigenvalues of L_α and phase transition for each α

Community Detection: improved approach

Generalized Laplacian: $L_\alpha = D^{-\alpha} A D^{-\alpha}$

Theoretical Results:

- Support of the eigenvalues of L_α and phase transition for each α
- "optimal" α depends on distribution of q_i but can be estimated

Community Detection: improved approach

Generalized Laplacian: $L_\alpha = D^{-\alpha} A D^{-\alpha}$

Theoretical Results:

- Support of the eigenvalues of L_α and phase transition for each α
- "optimal" α depends on distribution of q_i but can be estimated
- eigenvectors of L_α must be normalized by $D^{\alpha-1}$.

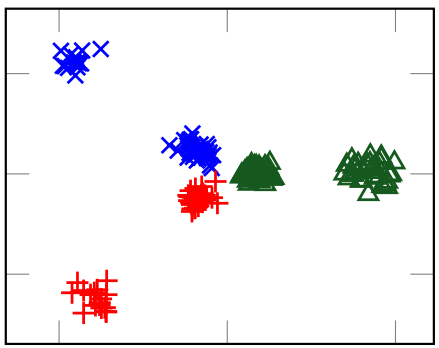
Community Detection: improved approach

Generalized Laplacian: $L_\alpha = D^{-\alpha} A D^{-\alpha}$

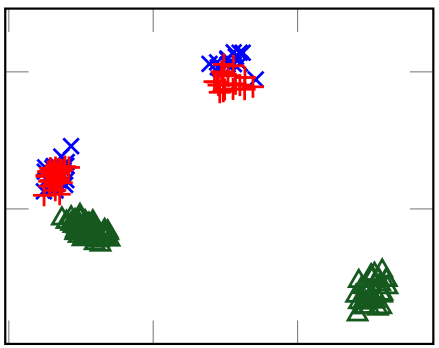
Theoretical Results:

- Support of the eigenvalues of L_α and phase transition for each α
- "optimal" α depends on distribution of q_i but can be estimated
- eigenvectors of L_α must be normalized by $D^{\alpha-1}$.

Performance in a synthetic scenario:



(Modularity A)



(Bethe Hessian $D - rA$)

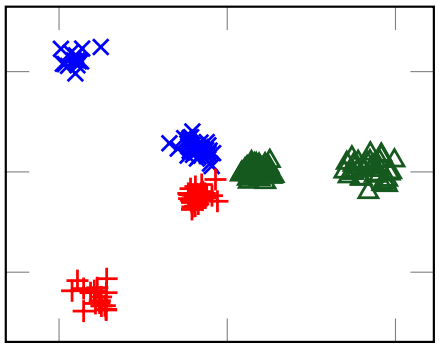
Community Detection: improved approach

Generalized Laplacian: $L_\alpha = D^{-\alpha} A D^{-\alpha}$

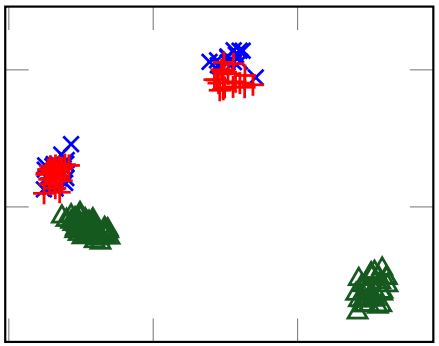
Theoretical Results:

- Support of the eigenvalues of L_α and phase transition for each α
- "optimal" α depends on distribution of q_i but can be estimated
- eigenvectors of L_α must be normalized by $D^{\alpha-1}$.

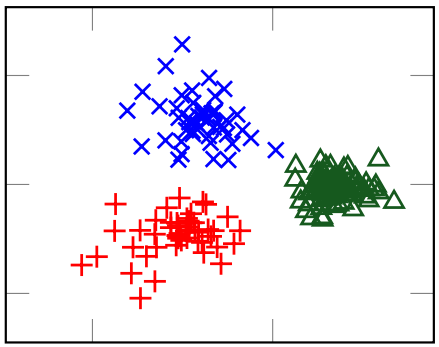
Performance in a synthetic scenario:



(Modularity A)



(Bethe Hessian $D - rA$)



(Algo with $\alpha = 1$)

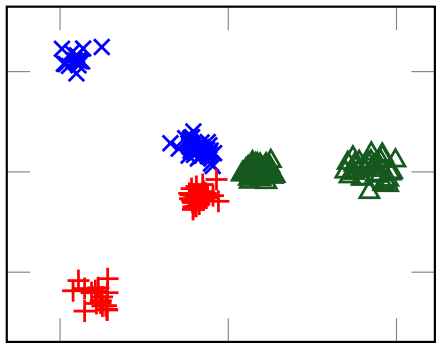
Community Detection: improved approach

Generalized Laplacian: $L_\alpha = D^{-\alpha} A D^{-\alpha}$

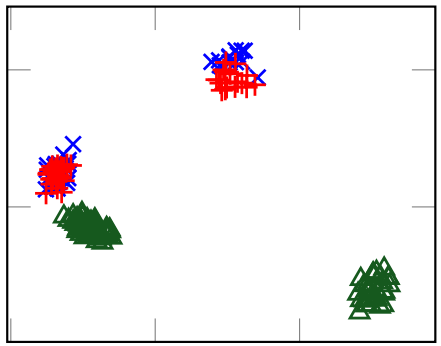
Theoretical Results:

- Support of the eigenvalues of L_α and phase transition for each α
- "optimal" α depends on distribution of q_i but can be estimated
- eigenvectors of L_α must be normalized by $D^{\alpha-1}$.

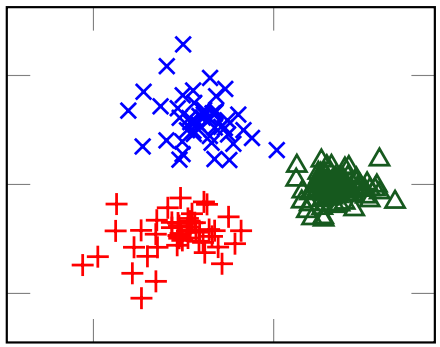
Performance in a synthetic scenario:



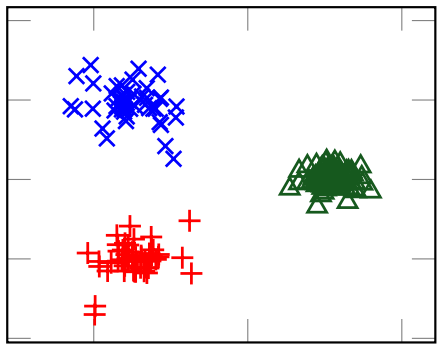
(Modularity A)



(Bethe Hessian $D - rA$)

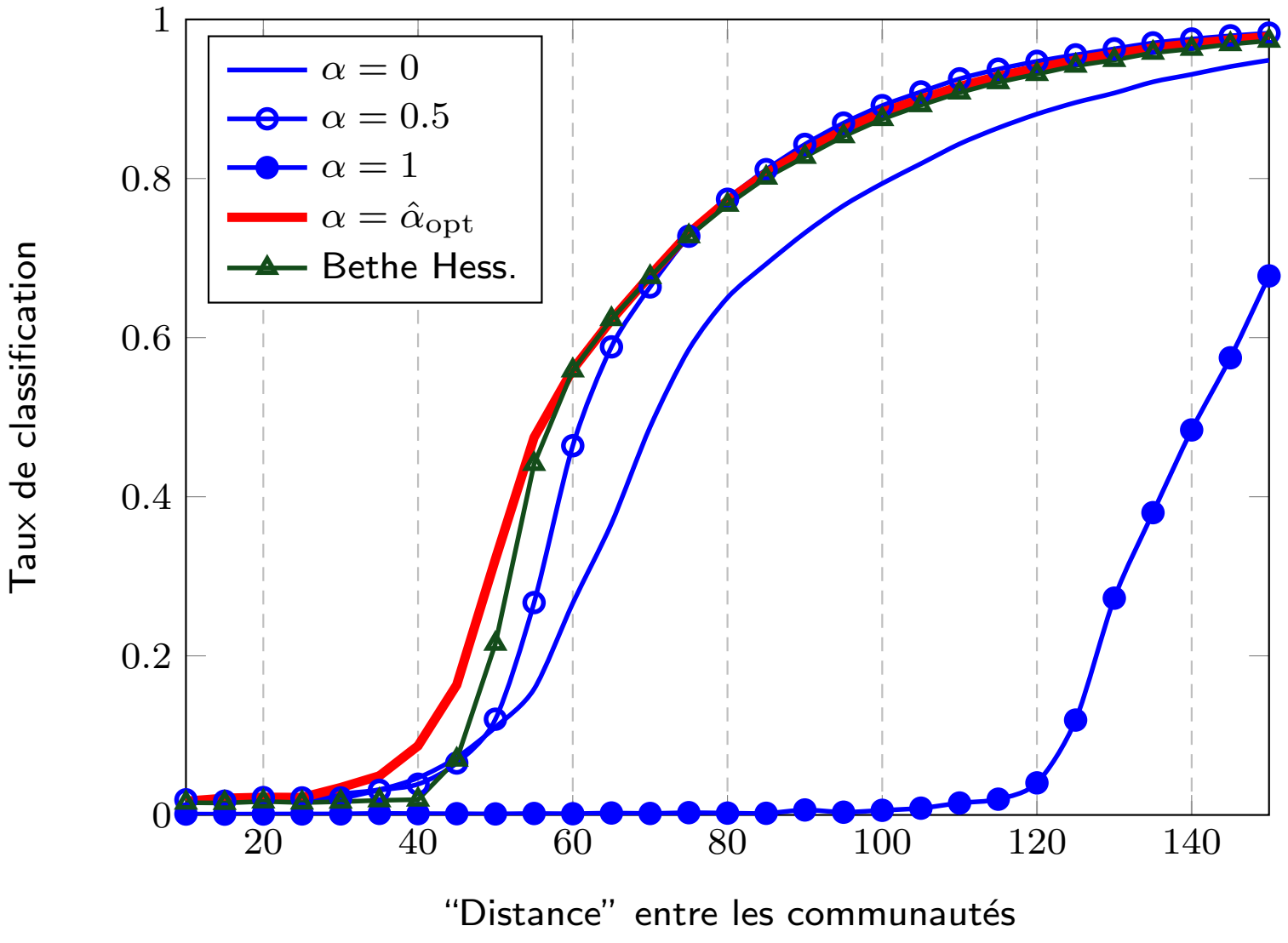


(Algo with $\alpha = 1$)



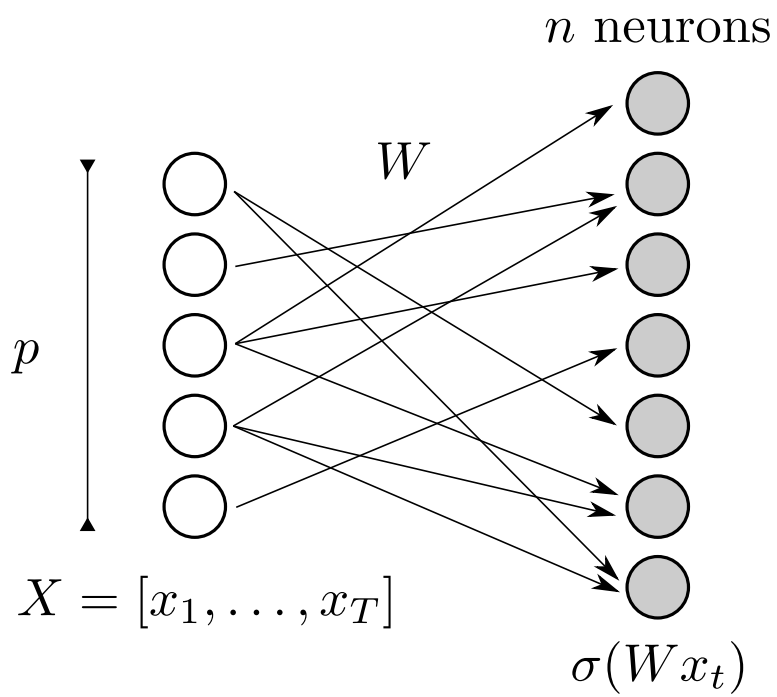
(Algo with $\hat{\alpha}_{\text{opt}}$)

Détection de communauté: notre méthode

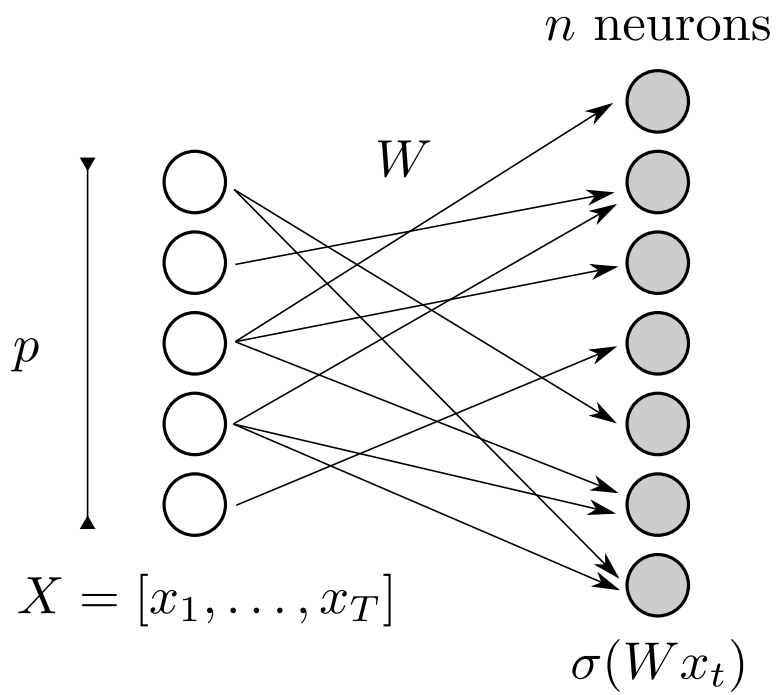


Outline

Random feature maps and extreme learning machines



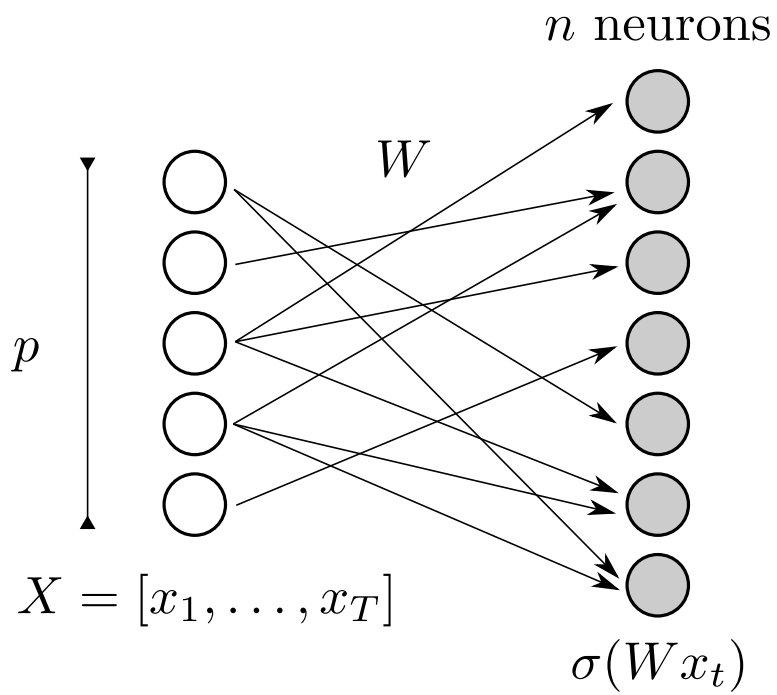
Random feature maps and extreme learning machines



Results:

- Limiting spectral analysis $(n, p, T \to \infty)$ of $\frac{1}{T}\Sigma\Sigma^{\mathsf{T}}$, $\Sigma_{ij} = \sigma(w_i^{\mathsf{T}} x_j)$.

Random feature maps and extreme learning machines

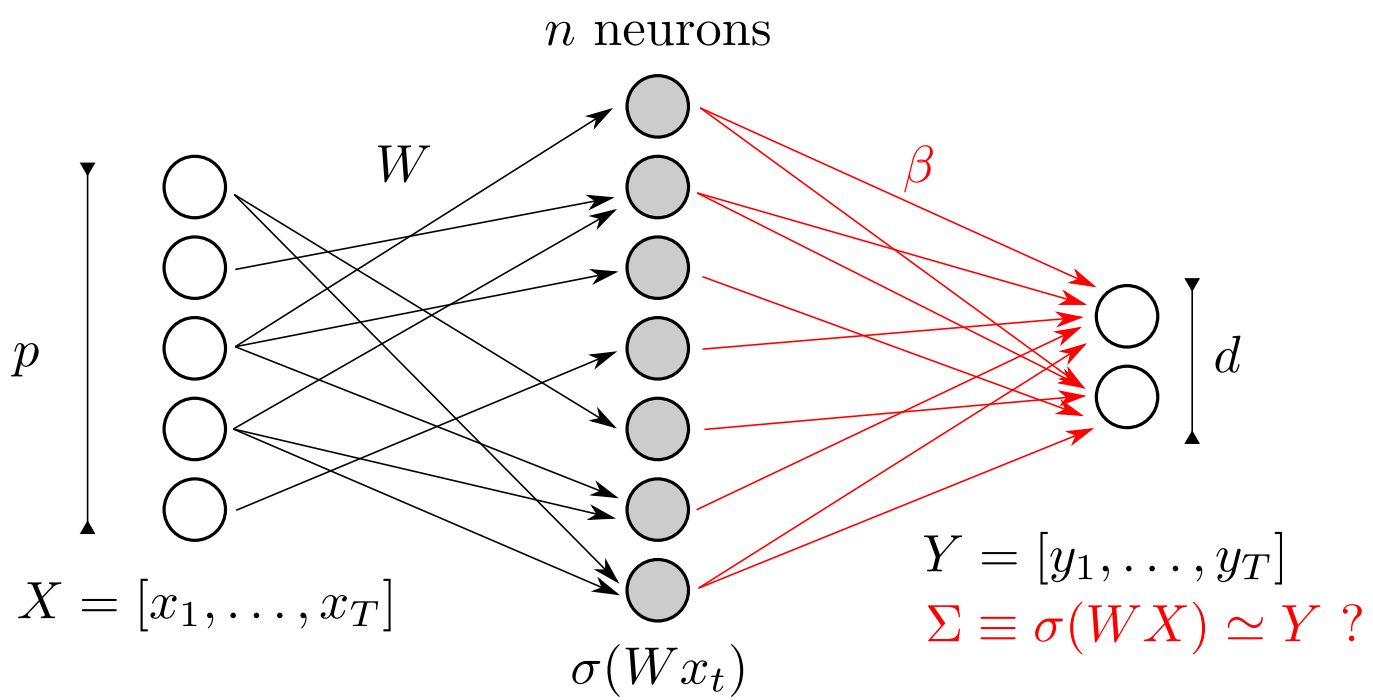


Results:

- Limiting spectral analysis $(n, p, T \to \infty)$ of $\frac{1}{T}\Sigma\Sigma^{\mathsf{T}}$, $\Sigma_{ij} = \sigma(w_i^{\mathsf{T}} x_j)$.

Concentration of measure phenomenon to break non-linearity in Σ.

Random feature maps and extreme learning machines

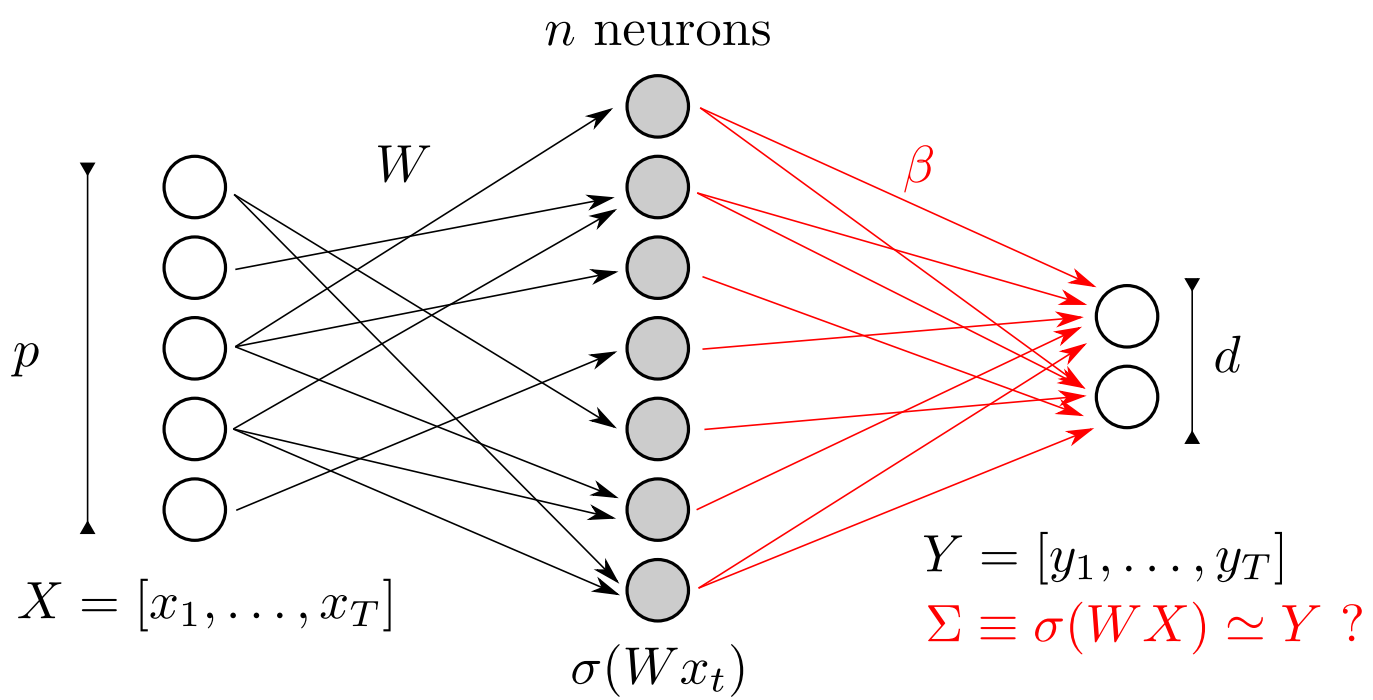


Results:

- Limiting spectral analysis ($n, p, T \to \infty$) of $\frac{1}{T}\Sigma\Sigma^{\mathsf{T}}$, $\Sigma_{ij} = \sigma(w_i^{\mathsf{T}} x_j)$.

Concentration of measure phenomenon to break non-linearity in Σ.

Random feature maps and extreme learning machines



Results:

- Limiting spectral analysis ($n, p, T \to \infty$) of $\frac{1}{T}\Sigma\Sigma^{\mathsf{T}}$, $\Sigma_{ij} = \sigma(w_i^{\mathsf{T}} x_j)$.

  Concentration of measure phenomenon to break non-linearity in Σ.

- Limiting error (MSE) in training (E_{train}) and testing (E_{test}).

(here $\beta = \frac{1}{T}\Sigma(\frac{1}{T}\Sigma^{\mathsf{T}}\Sigma + \gamma I_T)^{-1} Y$ for learning based on X, Y)

Simulations for MNIST with different $\sigma(\cdot)$

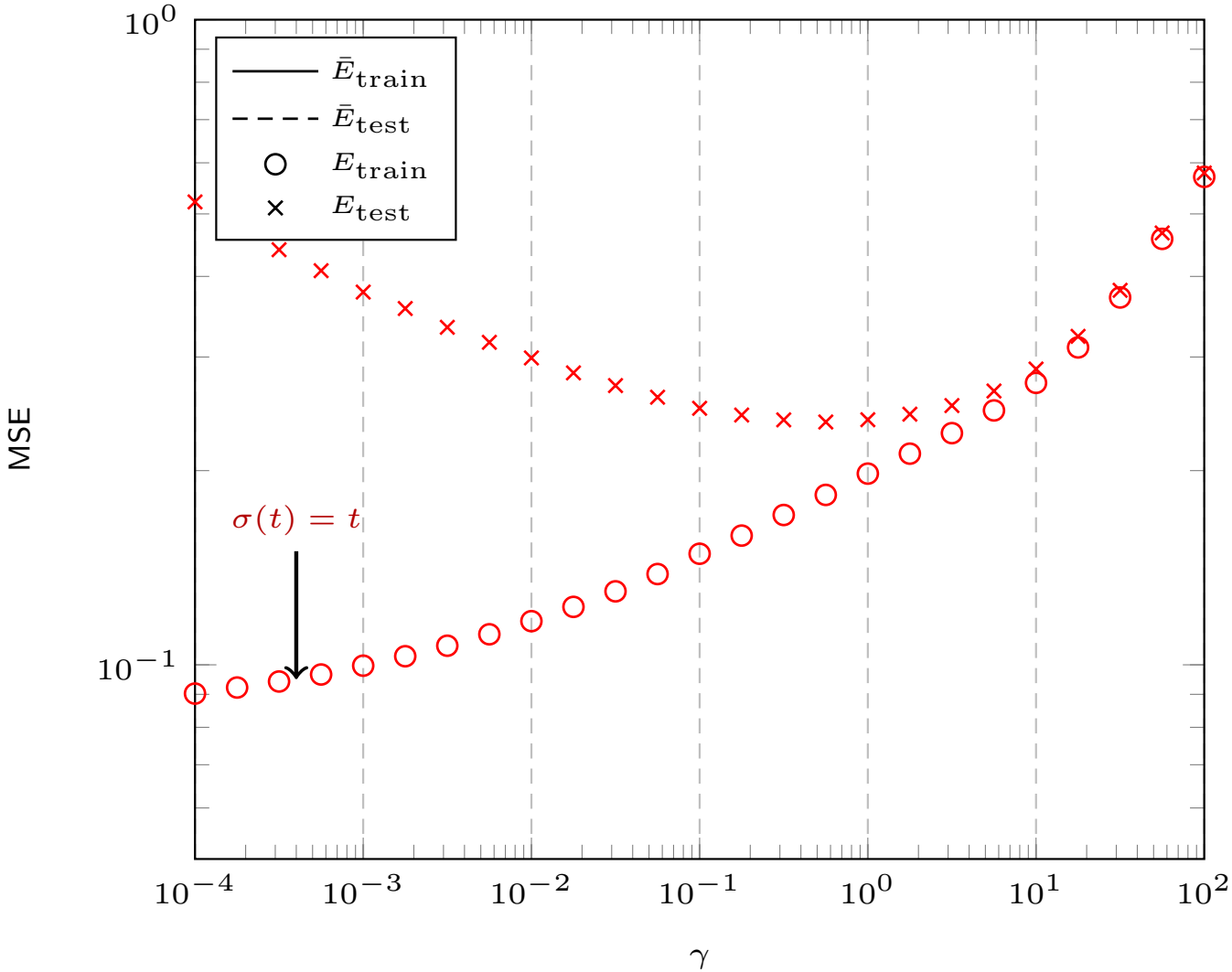


Figure: MSE performance, as a function of γ, two-class MNIST (7,9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations for MNIST with different $\sigma(\cdot)$

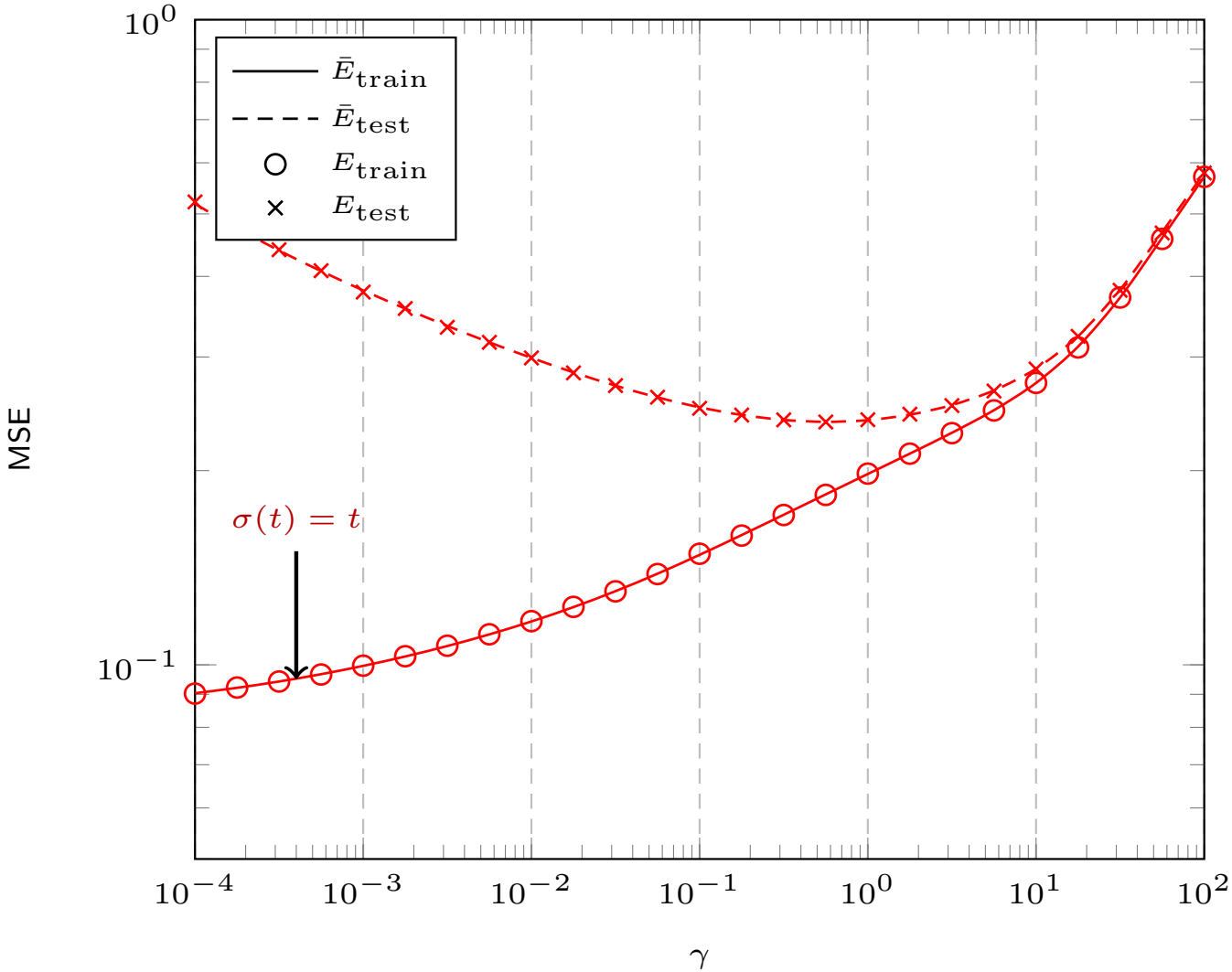


Figure: MSE performance, as a function of γ, two-class MNIST (7,9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations for MNIST with different $\sigma(\cdot)$

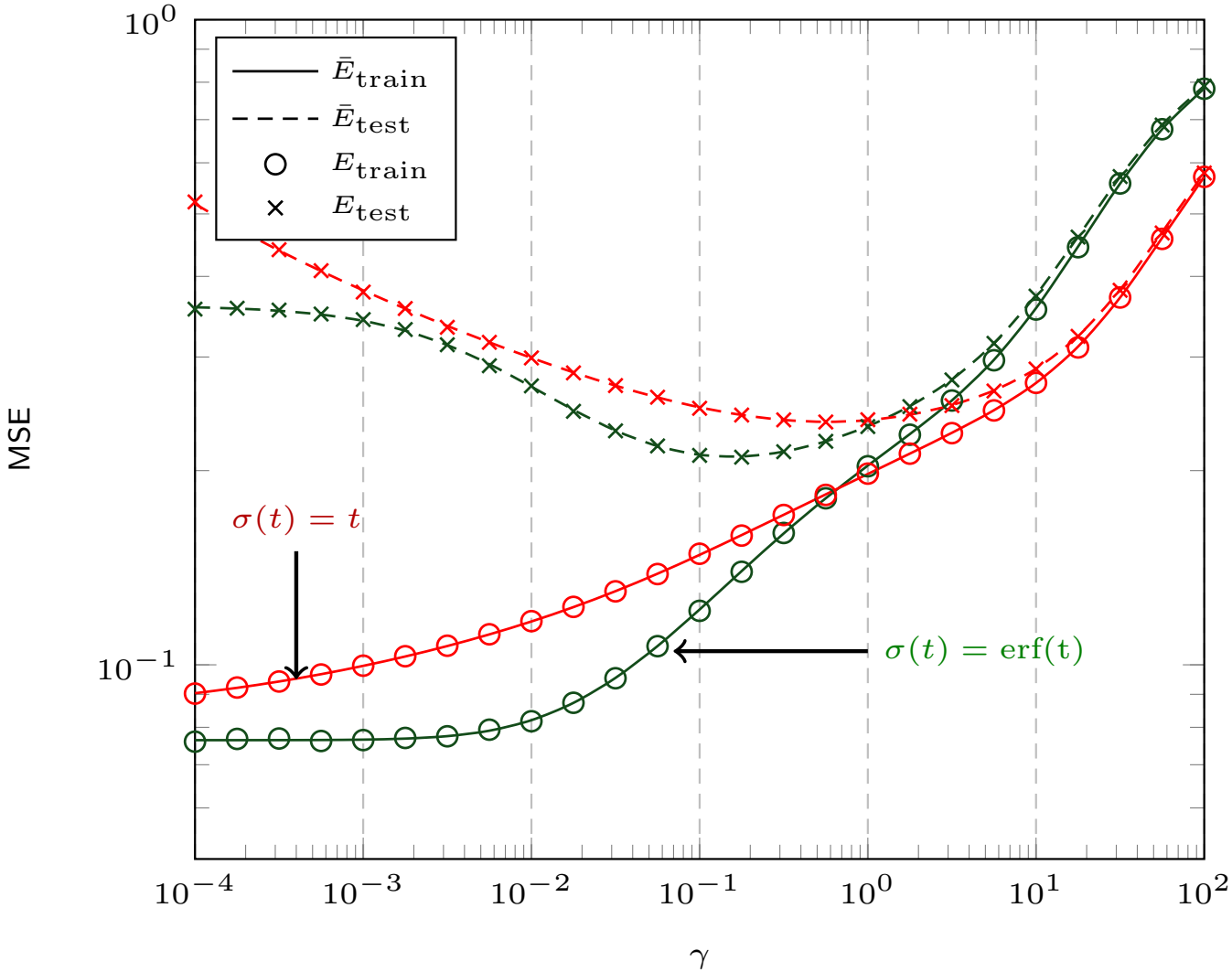


Figure: MSE performance, as a function of γ, two-class MNIST (7,9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations for MNIST with different $\sigma(\cdot)$

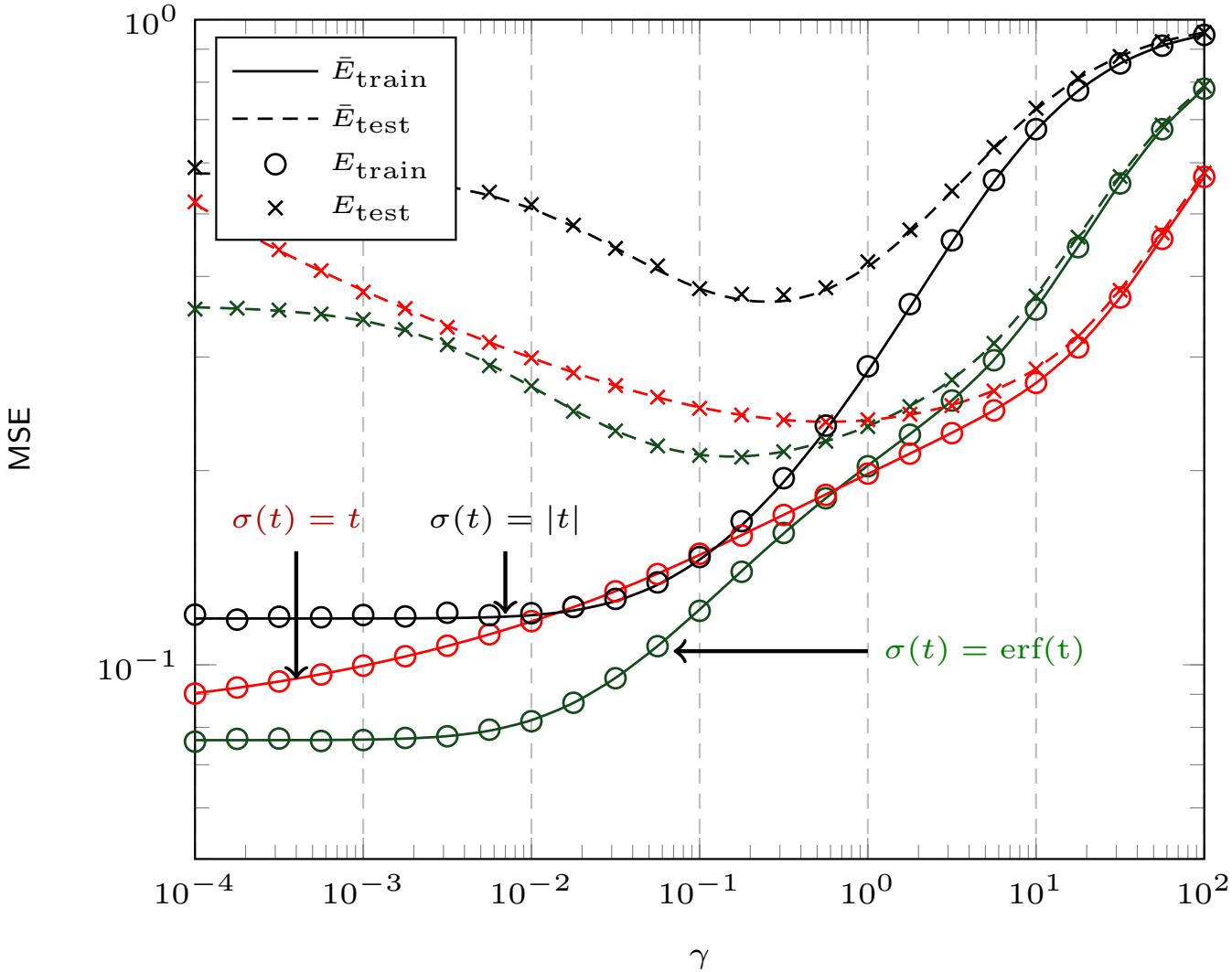


Figure: MSE performance, as a function of γ, two-class MNIST (7,9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations for MNIST with different $\sigma(\cdot)$

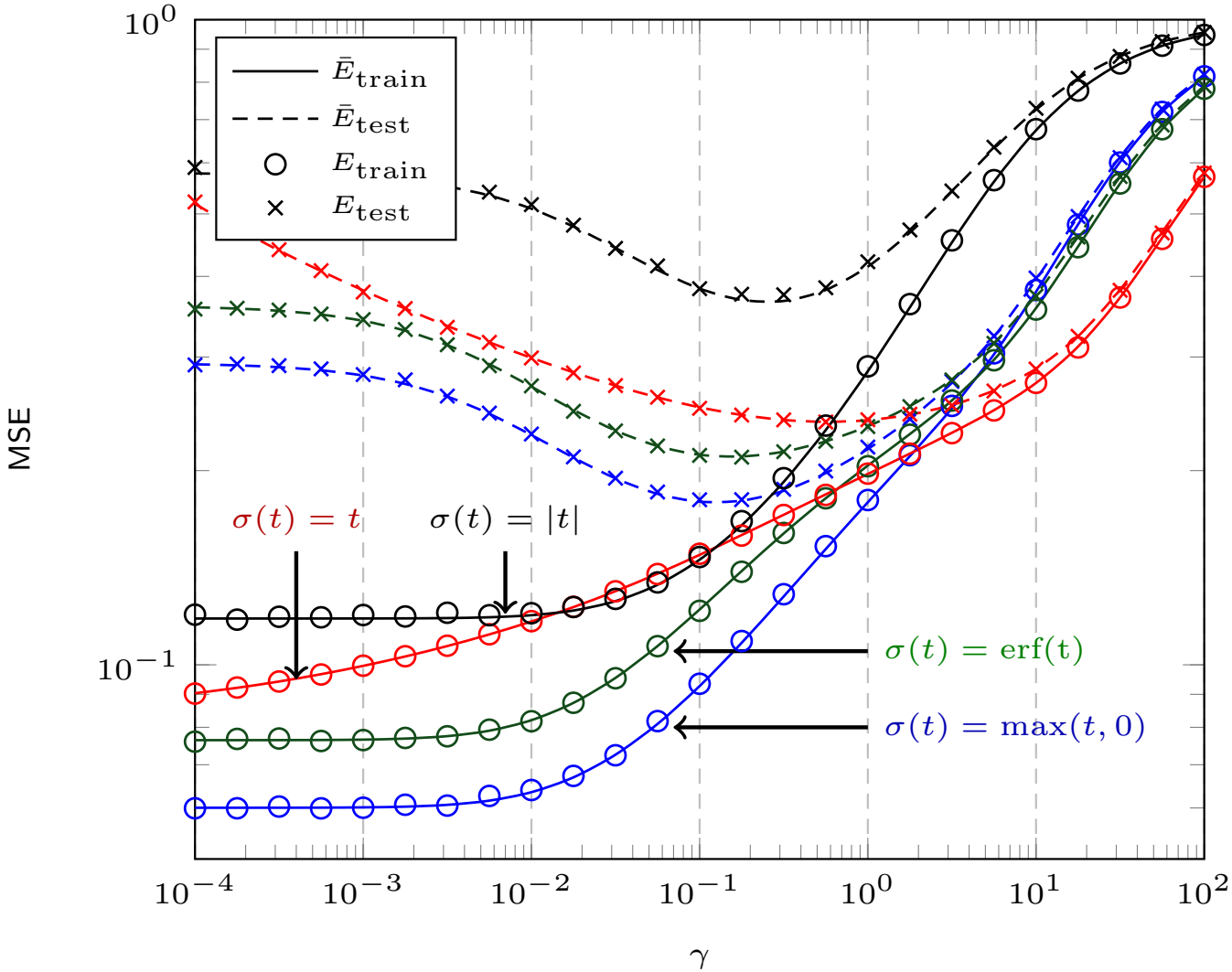


Figure: MSE performance, as a function of γ, two-class MNIST (7,9), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Outline

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions
  - improved methods, particularly improved "PageRank"
  - asymptotic performances, no phase transition

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions
  - improved methods, particularly improved "PageRank"
  - asymptotic performances, no phase transition
- LS-SVM supervised classification.
  - Similar results

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions
  - improved methods, particularly improved "PageRank"
  - asymptotic performances, no phase transition
- LS-SVM supervised classification.
  - Similar results
- Spectral community detection on large dimensional **heterogeneous** graphs
  - inconsistent standard methods

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions
  - improved methods, particularly improved "PageRank"
  - asymptotic performances, no phase transition
- LS-SVM supervised classification.
  - Similar results
- Spectral community detection on large dimensional **heterogeneous** graphs
  - inconsistent standard methods
  - new algorithm based on generalized Laplacian $D^{-\alpha} A D^{-\alpha}$

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions
  - improved methods, particularly improved "PageRank"
  - asymptotic performances, no phase transition
- LS-SVM supervised classification.
  - Similar results
- Spectral community detection on large dimensional **heterogeneous** graphs
  - inconsistent standard methods
  - new algorithm based on generalized Laplacian $D^{-\alpha} A D^{-\alpha}$
- Neural nets
  - accounts for non-linearity (doesn't linearize or Taylor expand)

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions
  - improved methods, particularly improved "PageRank"
  - asymptotic performances, no phase transition
- LS-SVM supervised classification.
  - Similar results
- Spectral community detection on large dimensional **heterogeneous** graphs
  - inconsistent standard methods
  - new algorithm based on generalized Laplacian $D^{-\alpha} A D^{-\alpha}$
- Neural nets
  - accounts for non-linearity (doesn't linearize or Taylor expand)
  - performances of random non-linear feedforward networks (extreme learning machines)
  - performances of random linear recursive networks (echo-state nets).

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions
  - improved methods, particularly improved "PageRank"
  - asymptotic performances, no phase transition
- LS-SVM supervised classification.
  - Similar results
- Spectral community detection on large dimensional **heterogeneous** graphs
  - inconsistent standard methods
  - new algorithm based on generalized Laplacian $D^{-\alpha} A D^{-\alpha}$
- Neural nets
  - accounts for non-linearity (doesn't linearize or Taylor expand)
  - performances of random non-linear feedforward networks (extreme learning machines)
  - performances of random linear recursive networks (echo-state nets).
- Random matrices and sparsity
  - sparse PCA with kernels

Summary of Main Contributions

List of articles, domains, results.

- Kernel spectral clustering
  - importance of $f(\tau), f'(\tau), f''(\tau)$ (more than "bandwidth")
  - surprising scenario for $f'(\tau) = 0$
  - asymptotic performances, phase transitions
- Semi-supervised learning approaches, graphs
  - intuition breaks in large dimensions
  - improved methods, particularly improved "PageRank"
  - asymptotic performances, no phase transition
- LS-SVM supervised classification.
  - Similar results
- Spectral community detection on large dimensional **heterogeneous** graphs
  - inconsistent standard methods
  - new algorithm based on generalized Laplacian $D^{-\alpha} A D^{-\alpha}$
- Neural nets
  - accounts for non-linearity (doesn't linearize or Taylor expand)
  - performances of random non-linear feedforward networks (extreme learning machines)
  - performances of random linear recursive networks (echo-state nets).
- Random matrices and sparsity
  - sparse PCA with kernels
- Large dimensional robust statistics
  - asymptotic equivalent of robust covariance estimators
  - numerous applications (finance, array processing, biostats)

Related References I

Kernels, classification, random projections, neural nets


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


A. Kammoun, R. Couillet, "Subspace Kernel Clustering of Large Dimensional Data", (soumis) Electronic Journal of Statistics, 2017.


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


X. Mai, R. Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data", (soumis) Journal of Machine Learning Research, 2017.


H. Tiomoko Ali, R. Couillet, "Spectral community detection in heterogeneous large networks", (soumis) Journal of Machine Learning Research, 2016.


C. Louart, Z. Liao, R. Couillet, "A Random Matrix Approach to Neural Networks", (à paraˆitre) Annals of Applied Probability, 2017.


R. Couillet, G. Wainrib, H. Sevi, H. Tiomoko Ali, "The asymptotic performance of linear echo state neural networks", Journal of Machine Learning Research, vol. 17, no. 178, pp. 1-35, 2016.

Related References II

Statistics, robust estimation, compressive sensing, signal


R. Couillet, M. McKay, "Optimal block-sparse PCA for high dimensional correlated samples", (submitted to) IEEE Trans. on Signal Processing, 2017.


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.


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.


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, 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.


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


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.

End

Thank You.