A Random Matrix Framework for BigData Machine Learning
(Groupe Deep Learning, DigiCosme)

Romain COUILLET

CentraleSupélec, France

June, 2017

Outline

Outline

Outline

Context

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

Context

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

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

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

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

Context

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

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

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

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

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

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

  or equivalently, in spectral norm

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

Context

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

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

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

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

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

  or equivalently, in spectral norm

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

Random Matrix Regime

- No longer valid if $p, n \to \infty$ with $p/n \to c \in (0, \infty)$,

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

Context

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

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

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

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

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

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

  or equivalently, in spectral norm

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

Random Matrix Regime

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

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

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

The Marčenko–Pastur law

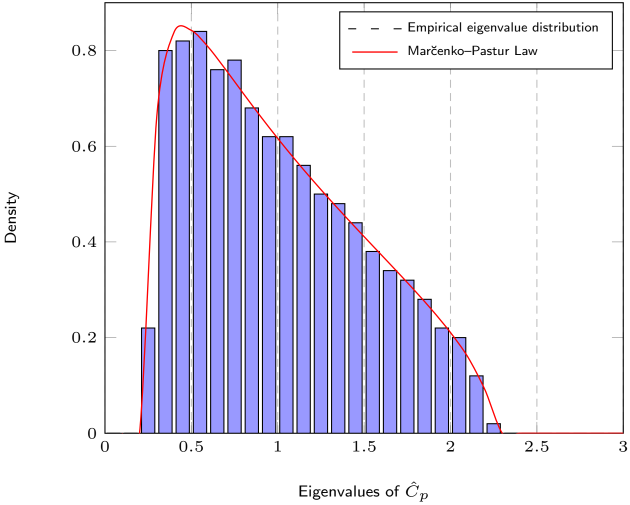


Figure: Histogram of the eigenvalues of $\hat{C}_p$ for $p = 500$, $n = 2000$, $C_p = I_p$.

Definition (Empirical Spectral Density)

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

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

Definition (Empirical Spectral Density)

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

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

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

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

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

weakly, where

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

Definition (Empirical Spectral Density)

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

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

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

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

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

weakly, where

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

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

The Marčenko–Pastur law

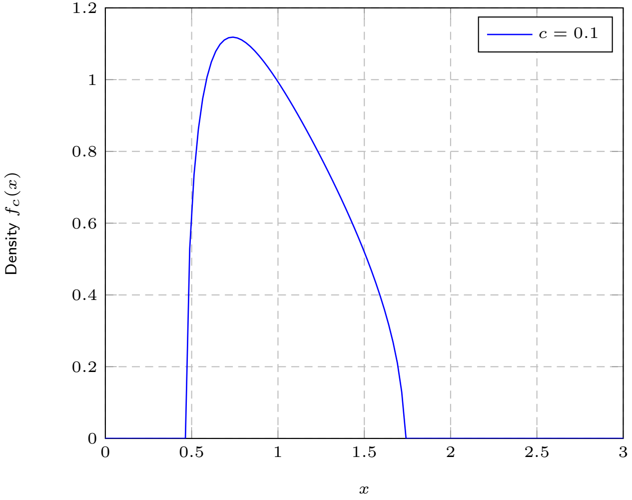


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

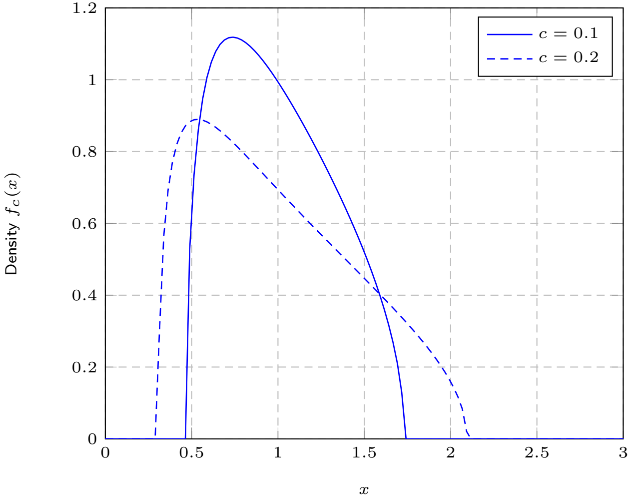


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

The Marčenko–Pastur law

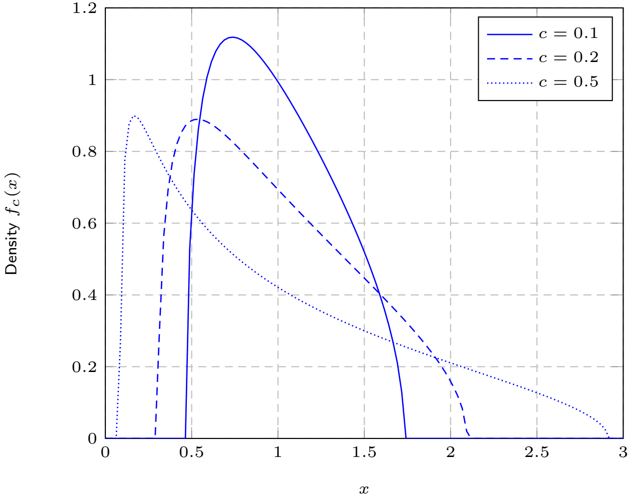


Figure: Marčenko-Pastur law for different limit ratios $c = \lim_{p\to\infty} p/n$.

Outline

If we break:

- **Small rank Perturbation**: $C_p = I_P + P$, P of low rank.

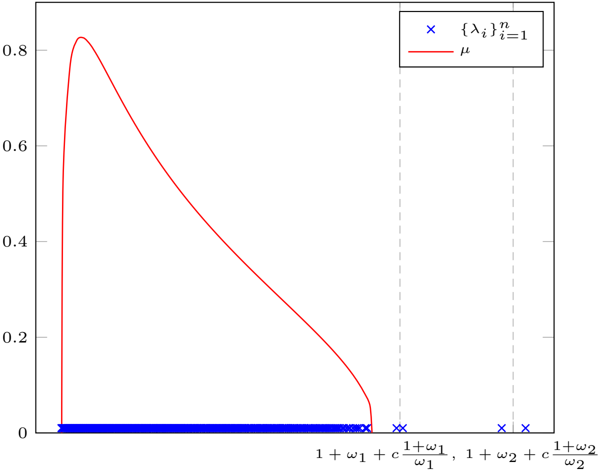


Figure: Eigenvalues of $\frac{1}{n}Y_pY_p^*$, $C_p = \mathrm{diag}(\underbrace{1,\ldots,1}_{p-4},2,2,3,3)$, $p=500$, $n=1500$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

In particular,

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

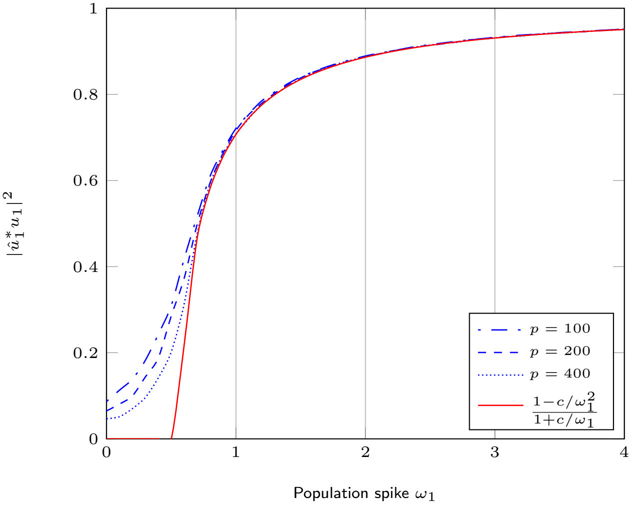


Figure: Simulated versus limiting $|\hat{u}_1^* u_1|^2$ for $Y_p = C_p^{\frac{1}{2}} X_p$, $C_p = I_p + \omega_1 u_1 u_1^*$, $p/n = 1/3$, varying ω_1.

Other Spiked Models

Similar results for multiple matrix models:

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

Outline

Outline

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 $U = [u_1, \ldots, u_\ell]$ with "dominant" eigenvalues

Reminder on Spectral Clustering Methods

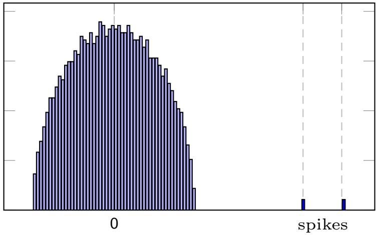
Context: Two-step classification of n objects based on similarity $A \in \mathbb{R}^{n\times n}$:

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

Reminder on Spectral Clustering Methods

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

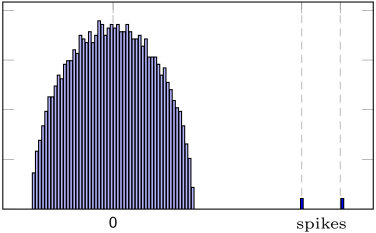
1. extraction of eigenvectors $U = [u_1, \ldots, u_\ell]$ with "dominant" eigenvalues
2. classification of n rows $U_{1,\cdot}, \ldots, U_{n,\cdot} \in \mathbb{R}^\ell$ using k-means/EM.



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 $U = [u_1, \ldots, u_\ell]$ with "dominant" eigenvalues
2. classification of n rows $U_{1,\cdot}, \ldots, U_{n,\cdot} \in \mathbb{R}^\ell$ using k-means/EM.

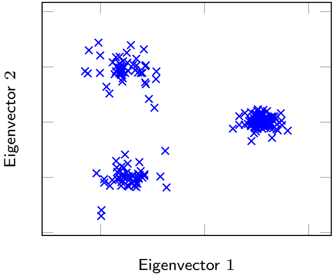


⇓ **Eigenvectors** ⇓
(in practice, shuffled)

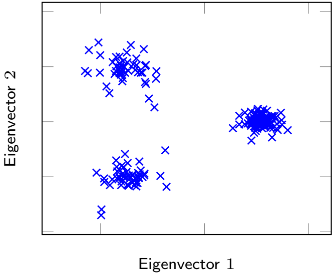
Reminder on Spectral Clustering Methods

Reminder on Spectral Clustering Methods

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



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



⇓
EM or k-means clustering.

Outline

Problem Statement

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

Kernel Spectral Clustering

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

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

Kernel Spectral Clustering

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

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

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

Kernel Spectral Clustering

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

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

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

Problem Statement

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

- Kernel spectral clustering based on kernel matrix

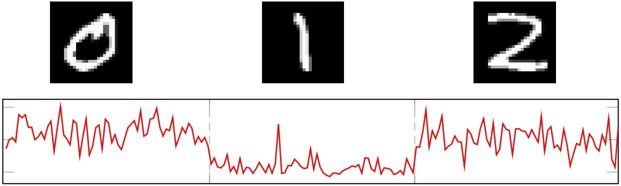
$$K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$$

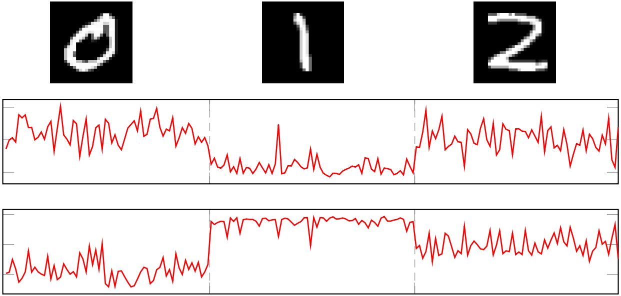
- Usually, $\kappa(x, y) = f(x^\mathsf{T} y)$ or $\kappa(x, y) = f(\|x - y\|^2)$
- Refinements:
  - instead of K, use $D - K$, $I_n - D^{-1}K$, $I_n - D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$, etc.
  - several steps algorithms: Ng–Jordan–Weiss, Shi–Malik, etc.

Intuition (from small dimensions)

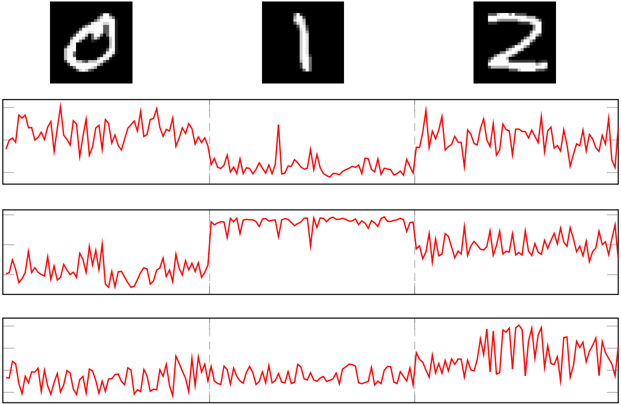
$$K = \begin{pmatrix} \kappa(x_i,x_j) \gg 1 & \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \ll 1 \\ \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \gg 1 & \kappa(x_i,x_j) \ll 1 \\ \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \ll 1 & \kappa(x_i,x_j) \gg 1 \end{pmatrix} \begin{matrix} \mathcal{C}_1 \\ \mathcal{C}_2 \\ \mathcal{C}_3 \end{matrix}$$

- K essentially low rank with class structure in eigenvectors.





Kernel Spectral Clustering



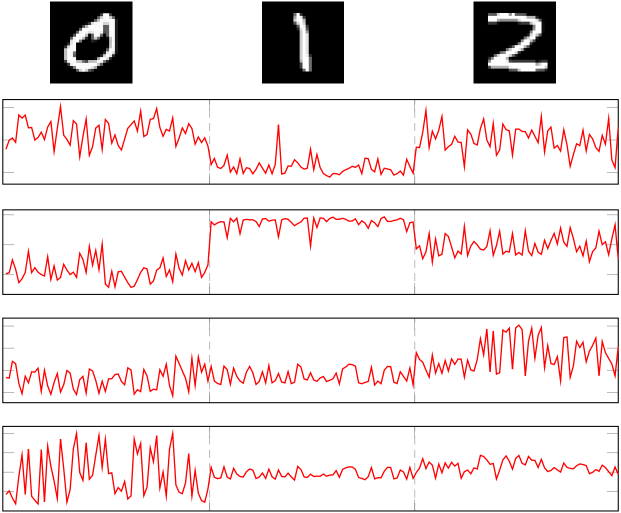


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data.

Model and Assumptions

Gaussian mixture model:

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

Gaussian mixture model:

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

Assumption (Convergence Rate)

As $n \to \infty$,

1. **Data scaling**: $\frac{p}{n} \to c_0 \in (0, \infty)$,
2. **Class scaling**: $\frac{n_a}{n} \to c_a \in (0, 1)$,
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*
$$\|C_a\| = O(1), \quad \operatorname{tr} C_a^\circ = O(\sqrt{p}), \quad \operatorname{tr} C_a^\circ C_b^\circ = O(p)$$

Gaussian mixture model:

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

Assumption (Convergence Rate)

As $n \to \infty$,

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

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

Remark: *For* 2 *classes, this is*

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

Kernel Matrix:

- Kernel matrix of interest:

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

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

Kernel Matrix:

- Kernel matrix of interest:

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

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

- We study the normalized Laplacian:

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

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

Random Matrix Equivalent

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

$$\boxed{\frac{1}{p} \|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau > 0.}$$

Random Matrix Equivalent

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

$$\frac{1}{p}\|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau > 0.$$

- Allows for Taylor expansion of K:

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

Random Matrix Equivalent

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

$$\boxed{\frac{1}{p}\|x_i - x_j\|^2 \xrightarrow{\text{a.s.}} \tau > 0.}$$

- Allows for Taylor expansion of K:

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

However not the (small dimension) intuitive behavior.

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

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

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

$$\hat{L} = -2 \frac{f'(\tau)}{f(\tau)} \left[\frac{1}{p} \Pi W^\mathsf{T} W \Pi + \frac{1}{p} J B J^\mathsf{T} + * \right]$$

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

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

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

$$\hat{L} = -2\frac{f'(\tau)}{f(\tau)} \left[\frac{1}{p} \Pi W^\mathsf{T} W \Pi + \frac{1}{p} J B J^\mathsf{T} + * \right]$$

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

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

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

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

Isolated eigenvalues: Gaussian inputs

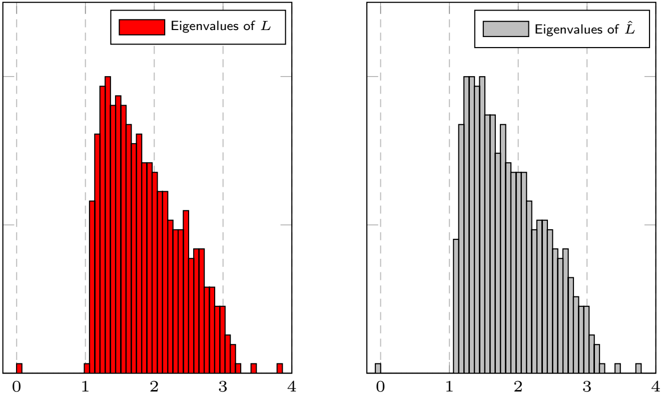


Figure: Eigenvalues of L and $\hat{L}$, $k = 3$, $p = 2048$, $n = 512$, $c_1 = c_2 = 1/4$, $c_3 = 1/2$, $[\mu_a]_j = 4\boldsymbol{\delta}_{aj}$, $C_a = (1 + 2(a-1)/\sqrt{p})I_p$, $f(x) = \exp(-x/2)$.

Theoretical Findings versus MNIST

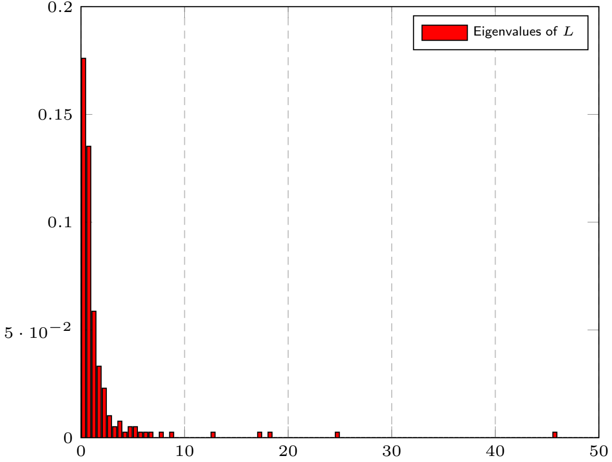


Figure: Eigenvalues of L (red) and (equivalent Gaussian model) $\hat{L}$ (white), MNIST data, $p = 784$, $n = 192$.

Theoretical Findings versus MNIST

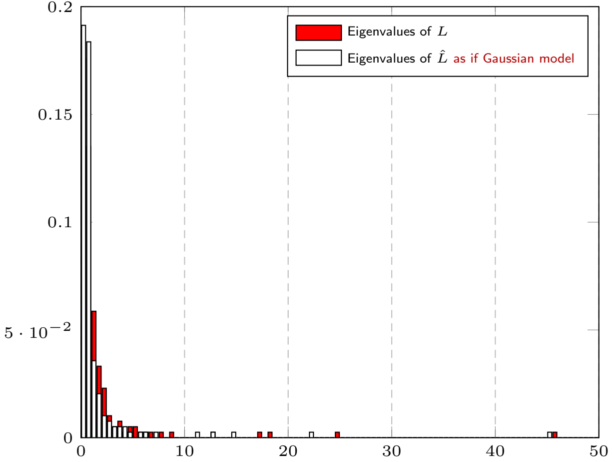


Figure: Eigenvalues of L (red) and (equivalent Gaussian model) $\hat{L}$ (white), MNIST data, $p = 784$, $n = 192$.

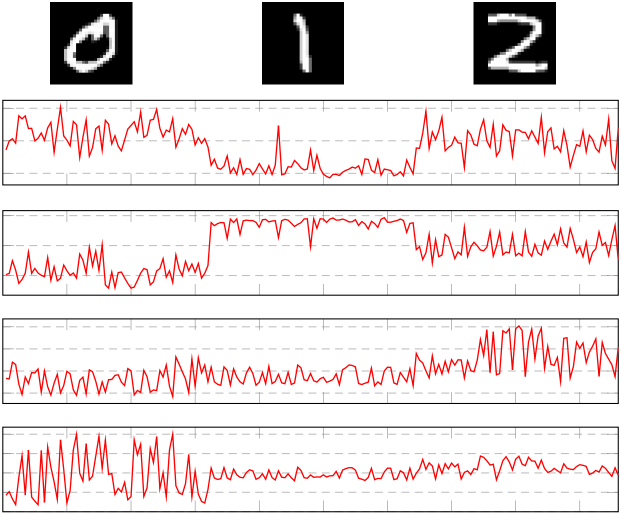


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theoretical Findings versus MNIST

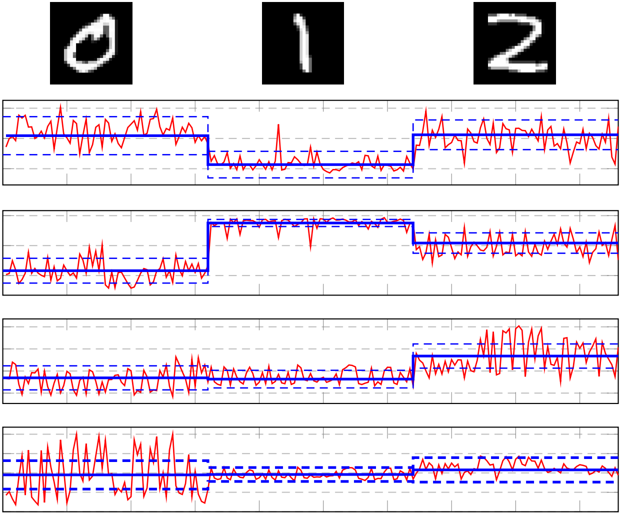


Figure: Leading four eigenvectors of $D^{-\frac{1}{2}} K D^{-\frac{1}{2}}$ for MNIST data (**red**) and theoretical findings (**blue**).

Theoretical Findings versus MNIST

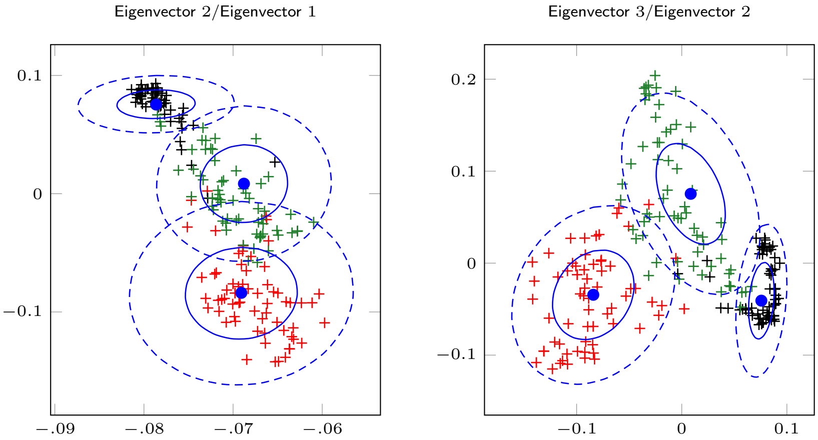


Figure: 2D representation of eigenvectors of L, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

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

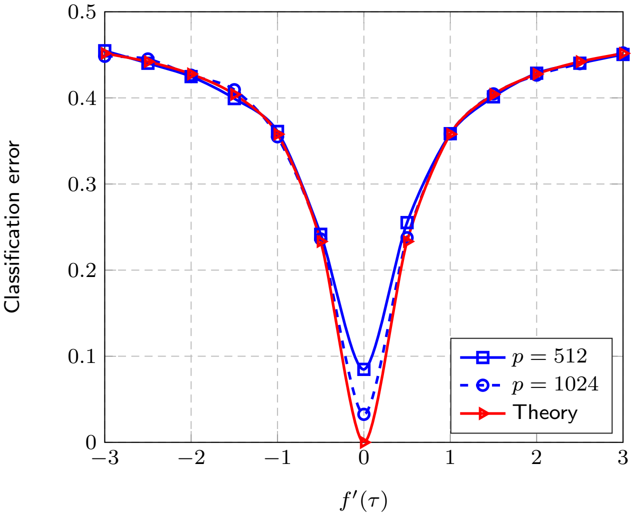


Figure: Classification performance, polynomial kernel with $f(\tau) = 4$, $f''(\tau) = 2$, $x_i \in \mathcal{N}(0, C_a)$, with $C_1 = I_p$, $[C_2]_{i,j} = .4^{|i-j|}$, $c_0 = \frac{1}{4}$.

Outline

Problem Statement

Context: Similar to clustering:

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

Problem Statement

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in k classes, with n_l labelled and n_u unlabelled data.
- Problem statement: give scores F_{ia} ($d_i = [K1_n]_i$)

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

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

Problem Statement

Context: Similar to clustering:

- Classify $x_1, \ldots, x_n \in \mathbb{R}^p$ in k classes, with n_l labelled and n_u unlabelled data.
- Problem statement: give scores F_{ia} ($d_i = [K1_n]_i$)

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

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

- **Solution**: for $F^{(u)} \in \mathbb{R}^{n_u \times k}$, $F^{(l)} \in \mathbb{R}^{n_l \times k}$ scores of unlabelled/labelled data,

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

where we naturally decompose

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

MNIST Data Example

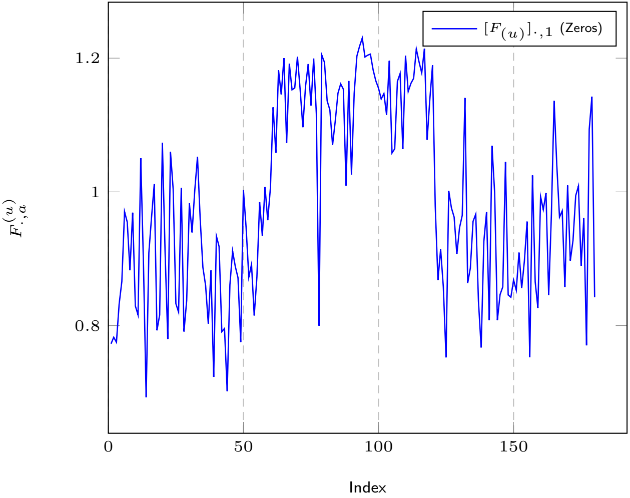


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

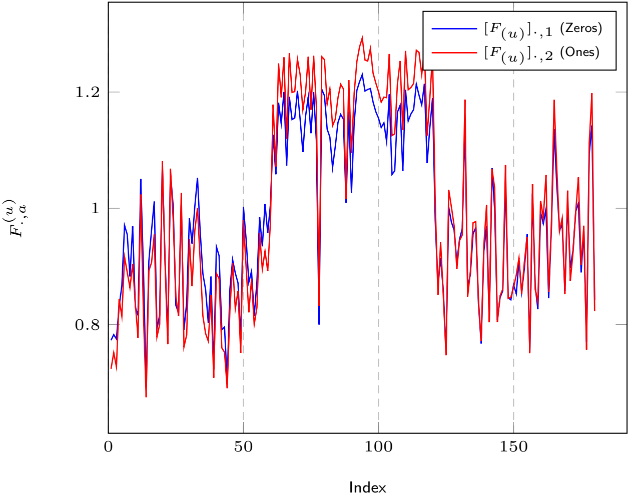


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

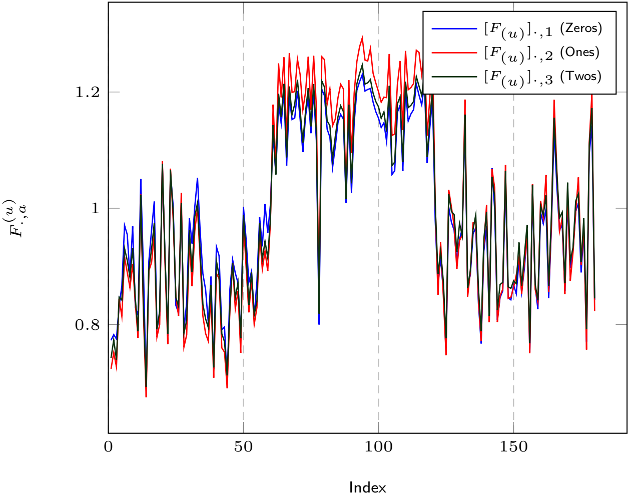


Figure: Vectors $[F^{(u)}]_{\cdot,a}$, $a = 1, 2, 3$, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

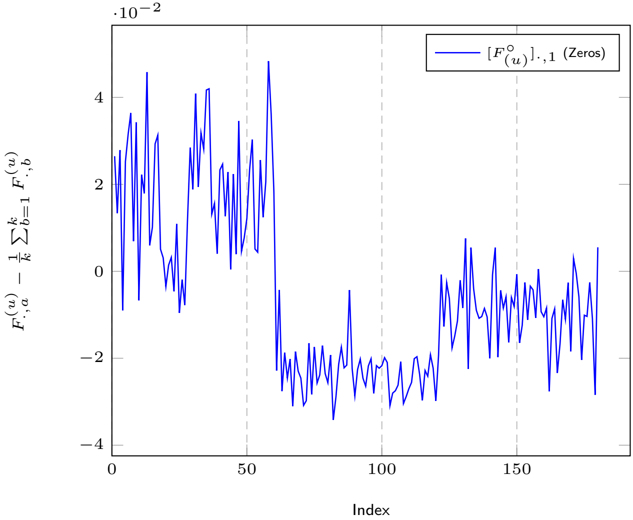


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k 1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

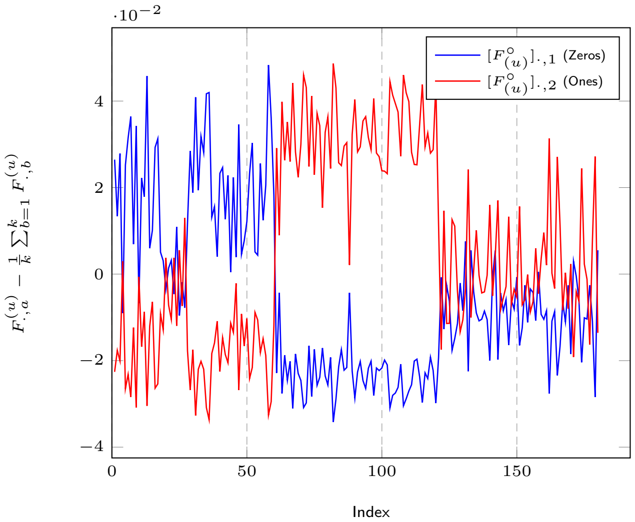


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k 1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

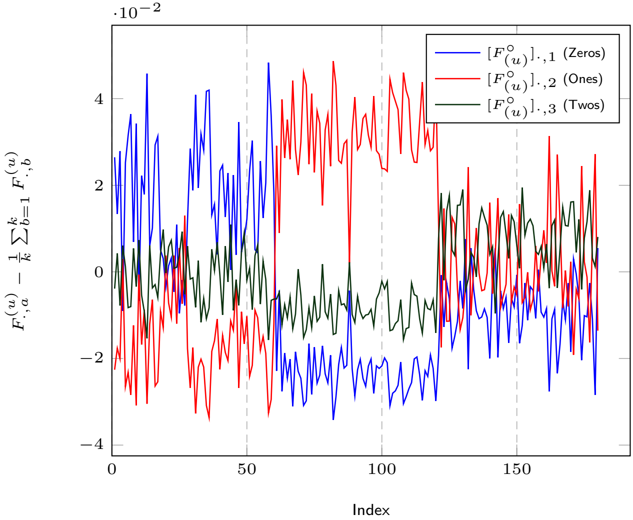


Figure: Centered Vectors $[F^{\circ}_{(u)}]_{\cdot,a} = [F_{(u)} - \frac{1}{k}F_{(u)}1_k1_k^{\mathsf{T}}]_{\cdot,a}$, 3-class MNIST data (zeros, ones, twos), $\alpha = 0$, $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Main Results

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

- In the first order,

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

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

Main Results

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

- In the first order,

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

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

- Consequences:

Main Results

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

- In the first order,

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

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

- Consequences:
  - Random non-informative bias v

Main Results

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

- In the first order,

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

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

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

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

Main Results

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

- In the first order,

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

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

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

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

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

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

Main Results

As a consequence of the remarks above, we take

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

and define

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

As a consequence of the remarks above, we take

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

and define

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

Theorem

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

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

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

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

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

with t, T, M as before, $\tilde{X}_a = X_a - \sum_{d=1}^k \frac{n_{l,d}}{n_l} X_d^\circ$ and B_b bias independent of a.

Corollary (Asymptotic Classification Error)

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

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

Corollary (Asymptotic Classification Error)

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

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

Some consequences:

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

MNIST Data Example

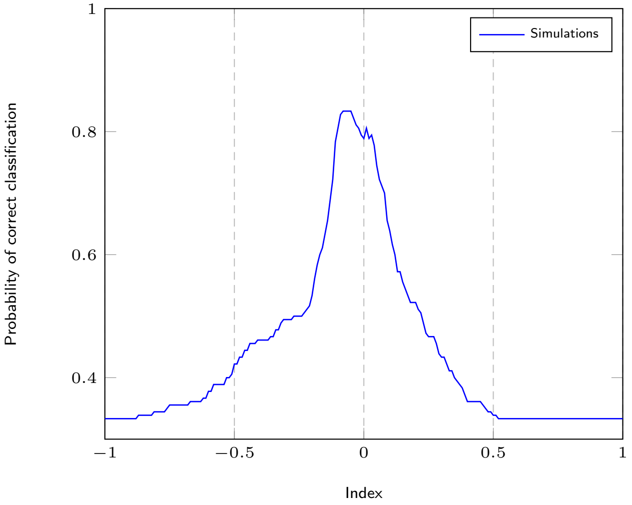


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

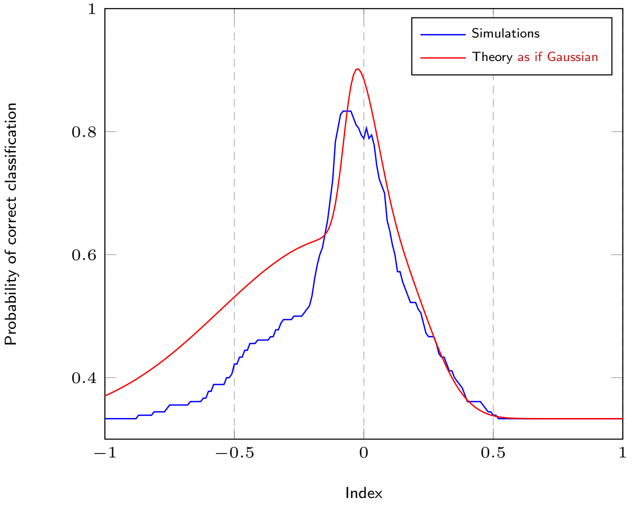


Figure: Performance as a function of α, for 3-class MNIST data (zeros, ones, twos), $n = 192$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

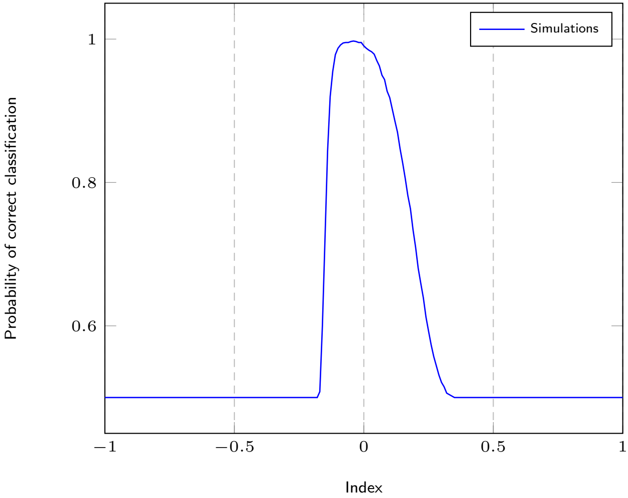


Figure: Performance as a function of α, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

MNIST Data Example

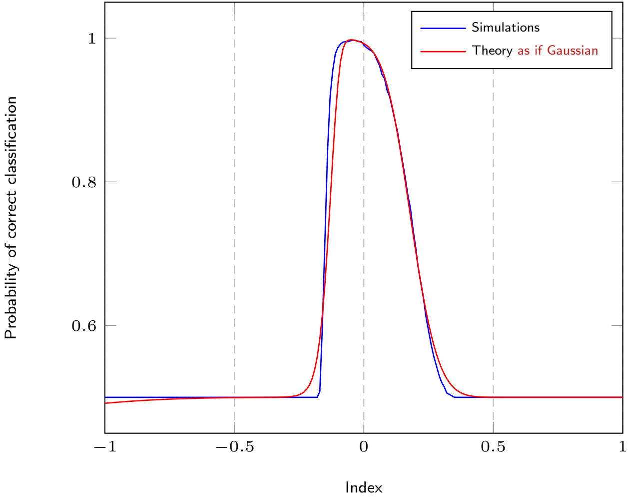
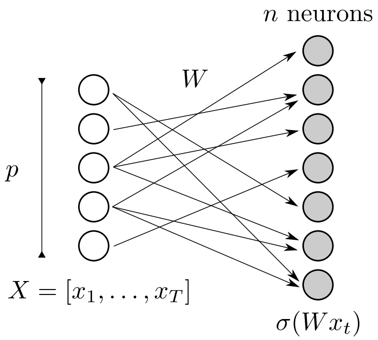


Figure: Performance as a function of α, for 2-class MNIST data (zeros, ones), $n = 1568$, $p = 784$, $n_l/n = 1/16$, Gaussian kernel.

Outline

Context: Random Feature Map

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



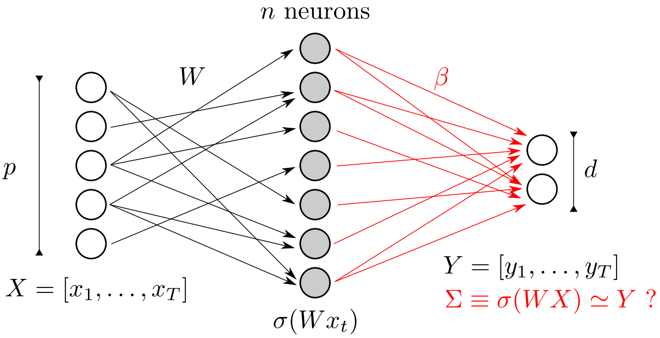
Random Feature Maps and Extreme Learning Machines

Context: Random Feature Map

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

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

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



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

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

- **Training MSE**:

$$E_{\rm train} = \frac{1}{T}\sum_{i=1}^{T} \|y_i - \beta^{\sf T}\sigma(Wx_i)\|^2 = \frac{1}{T}\|Y - \beta^{\sf T}\Sigma\|_F^2$$

with

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

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

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

- **Training MSE**:

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

with

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

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

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

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

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

Preliminary observations:

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

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

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

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

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

Preliminary observations:

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

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

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

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

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

Central object: resolvent $E[Q]$.

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

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

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

for some $C > 0$, where

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

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

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

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

Main Technical Result

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

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

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

for some $C > 0$, where

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

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

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

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

Proof arguments:

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

Main Technical Result

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

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

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

for some $C > 0$, where

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

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

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

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

Proof arguments:

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

Concentration of measure lemma: $\frac{1}{p} \sigma(w^\mathsf{T} X) A \sigma(X^\mathsf{T} w) \simeq \frac{1}{p} \operatorname{tr} \Phi A$

Main Technical Result

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

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

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

Main Technical Result

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

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

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

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

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

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

Theorem [Asymptotic $E_{\rm train}$]

For all $\varepsilon > 0$,

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

almost surely, where

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

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

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

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

Claim [Asymptotic E_{test}]

For all $\varepsilon > 0$,

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

almost surely, where

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

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

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

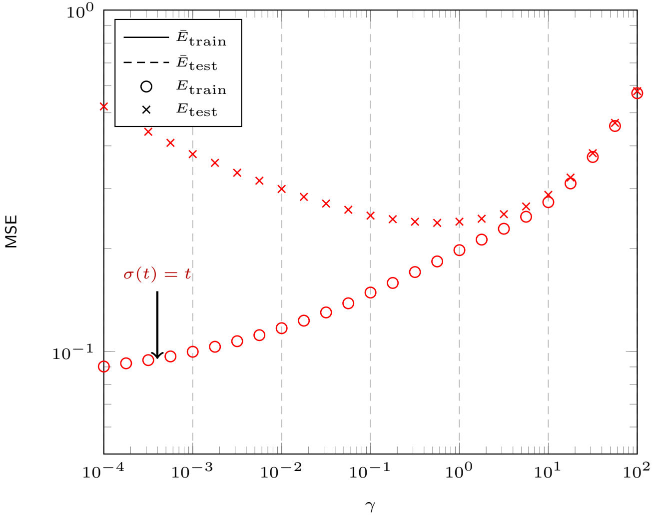


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

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

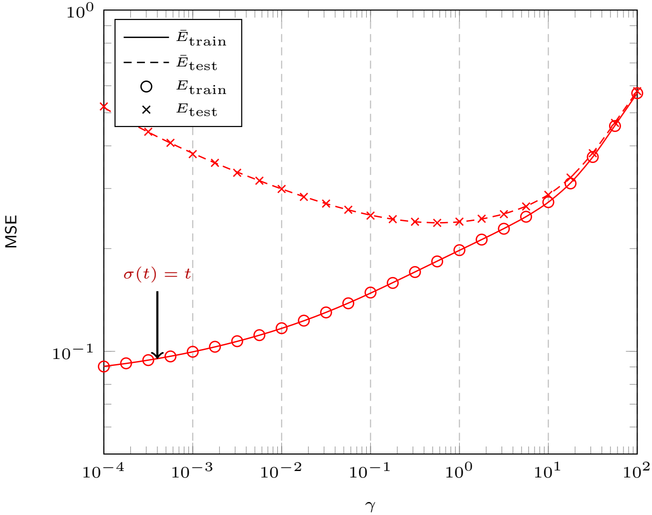


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

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

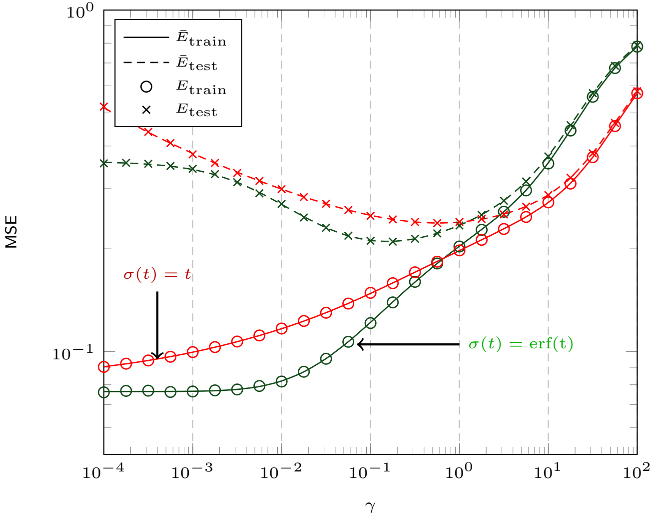


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

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

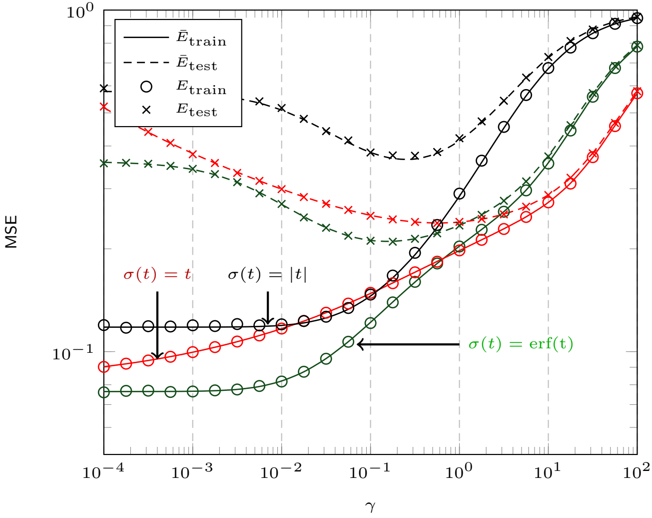


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

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

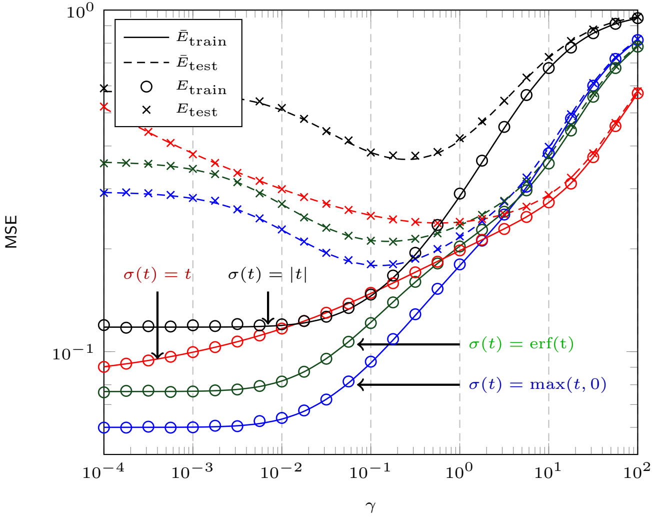


Figure: Neural network performance for Lipschitz continuous $\sigma(\cdot)$, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

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

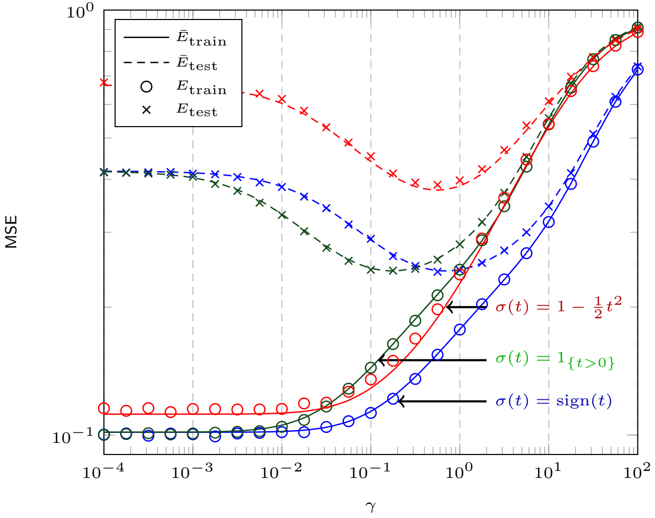


Figure: Neural network performance for $\sigma(\cdot)$ either discontinuous or non Lipschitz, as a function of γ, for 2-class MNIST data (sevens, nines), $n = 512$, $T = \hat{T} = 1024$, $p = 784$.

Simulations on “tuned” Gaussian mixture

Gaussian mixture classification

- $X = [X_1, X_2]$, with $\{X_1\}_i \sim \mathcal{N}(0, C_1)$, $\{X_2\}_i \sim \mathcal{N}(0, C_2)$, $\operatorname{tr} C_1 = \operatorname{tr} C_2$

Simulations on "tuned" Gaussian mixture

Gaussian mixture classification

- $X = [X_1, X_2]$, with $\{X_1\}_i \sim \mathcal{N}(0, C_1)$, $\{X_2\}_i \sim \mathcal{N}(0, C_2)$, $\operatorname{tr} C_1 = \operatorname{tr} C_2$
- We can prove that, for $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$ and $E[W_{ij}^k] = m_k$,

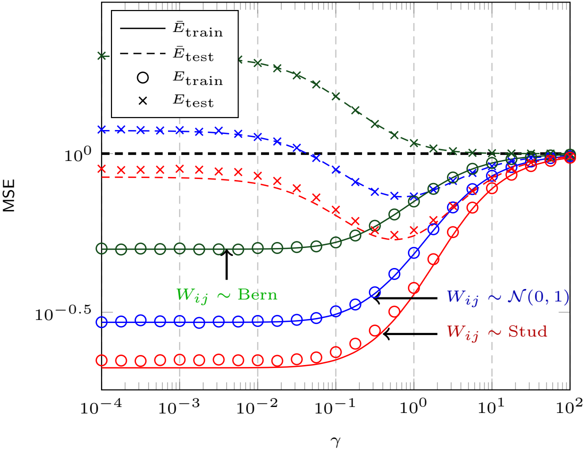
$\longrightarrow$ Classification only possible if $m_4 \neq m_2^2$

Simulations on “tuned” Gaussian mixture

Gaussian mixture classification

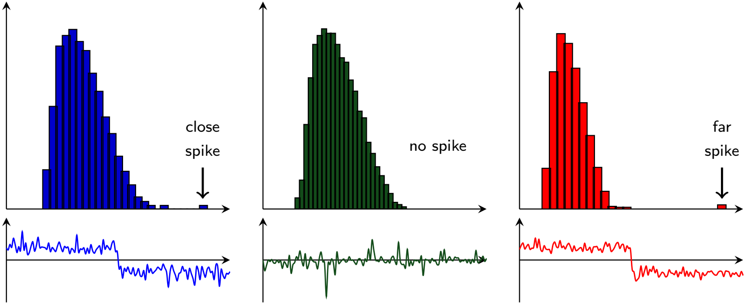
- $X = [X_1, X_2]$, with $\{X_1\}_i \sim \mathcal{N}(0, C_1)$, $\{X_2\}_i \sim \mathcal{N}(0, C_2)$, $\operatorname{tr} C_1 = \operatorname{tr} C_2$
- We can prove that, for $\sigma(t) = \zeta_2 t^2 + \zeta_1 t + \zeta_0$ and $E[W_{ij}^k] = m_k$,

$\longrightarrow$ Classification only possible if $m_4 \neq m_2^2$



Simulations on "tuned" Gaussian mixture

- Interpretation in eigenstructure of Φ: no information carried in dominant eigenmodes if $m_4 = m_2^2$.



$W_{ij} \sim \mathcal{N}(0,1)$ $W_{ij} \sim \text{Bern}$ $W_{ij} \sim \text{Stud}$

Outline

Summary of Results and Perspectives I

Random Neural Networks.

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

Deep Neural Networks (DNN).

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

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

Summary of Results and Perspectives II

References.

H. W. Lin, M. Tegmark, “Why does deep and cheap learning work so well?”, arXiv:1608.08225v2, 2016.

C. Williams, “Computation with infinite neural networks”, Neural Computation, 10(5), 1203-1216, 1998.

Herbert Jaeger. Short term memory in echo state networks. GMD-Forschungszentrum Informationstechnik, 2001.

Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew, “Extreme learning machine : theory and applications”, Neurocomputing, 70(1) :489501, 2006.

N. El Karoui, “Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond”, The Annals of Applied Probability, 19(6), 2362-2405, 2009.

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

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

Choromanska, Anna, et al. ”The Loss Surfaces of Multilayer Networks.” AISTATS. 2015.

Summary of Results and Perspectives III

Rahimi, Ali, and Benjamin Recht. "Random Features for Large-Scale Kernel Machines." NIPS. Vol. 3. No. 4. 2007.

Summary of Results and Perspectives I

Kernel methods.

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

Applications.

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

References.

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

- R. Couillet, F. Benaych-Georges, “Kernel Spectral Clustering of Large Dimensional Data”, Electronic Journal of Statistics, vol. 10, no. 1, pp. 1393-1454, 2016.
- R. Couillet, A. Kammoun, “Random Matrix Improved Subspace Clustering”, Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2016.
- Z. Liao, R. Couillet, "A Large Dimensional Analysis of Least Squares Support Vector Machines”, (submitted to) Journal of Machine Learning Research, 2017.
- X. Mai, R. Couillet, “The counterintuitive mechanism of graph-based semi-supervised learning in the big data regime”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17), New Orleans, USA, 2017.

Community detection.

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

Applications.

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

References.

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

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

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

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

Summary of Results and Perspectives I

Robust statistics.

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

Applications.

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

References.

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

Summary of Results and Perspectives II

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

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

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

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

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

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

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

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

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

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

Summary of Results and Perspectives I

Other works and ideas.

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

Applications.

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

References.

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

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

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

The End

Thank you.