# Spectral clustering and random matrices

Florent BENAYCH-GEORGES (a)
joint work with Romain COUILLET (b)

(a) Université Paris Descartes
(b) CentraleSupélec

May 4, 2016
Lille

# Clustering

**Goal** : cluster observations $x_1, \ldots, x_n$

$$x_1 \quad , \ldots , \quad x_n \quad \Longrightarrow \quad \mathcal{C}_1 = \{x_3, x_{18}, \ldots\}, \ldots, \mathcal{C}_k = \{x_1, x_{20}, \ldots\}$$

$$\downarrow \qquad\qquad \downarrow$$

$$\left.\begin{bmatrix} x_{1,1} \\ \vdots \\ x_{1,p} \end{bmatrix} \quad \begin{bmatrix} x_{n,1} \\ \vdots \\ x_{n,p} \end{bmatrix}\right\} p \text{ coordinates for each observation : } x_i \in \mathbb{R}^p$$

**Examples** : $k$-means, EM, hierarchical clustering

# Clustering

**Goal** : cluster observations $x_1, \ldots, x_n$ with maximum similarity intra classes and minimum similarity inter classes

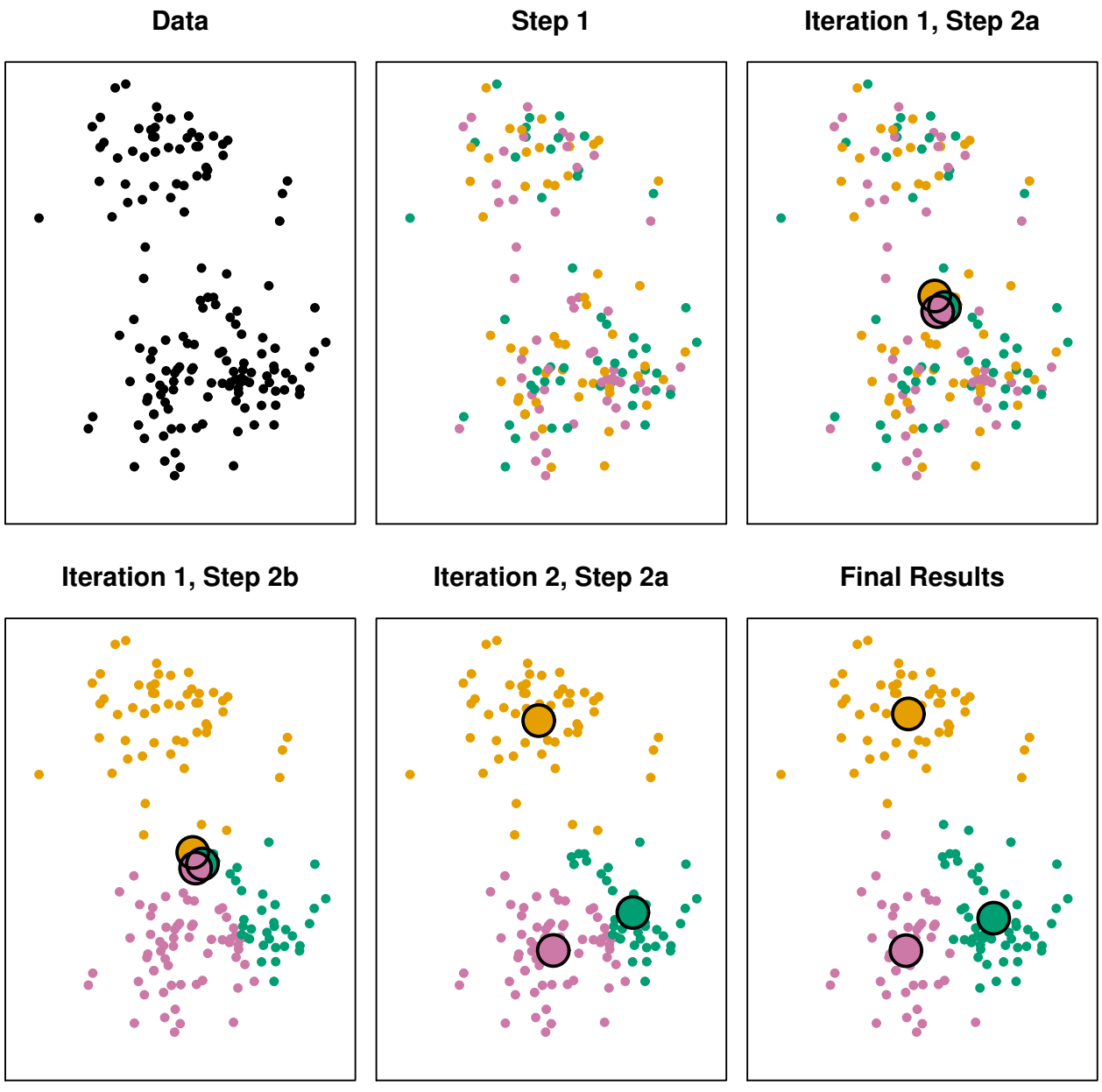
$$x_1 \quad , \ldots , \quad x_n \quad \Longrightarrow \quad \mathcal{C}_1 = \{x_3, x_{18}, \ldots\}, \ldots, \mathcal{C}_k = \{x_1, x_{20}, \ldots\}$$

$$\downarrow \qquad\qquad \downarrow$$

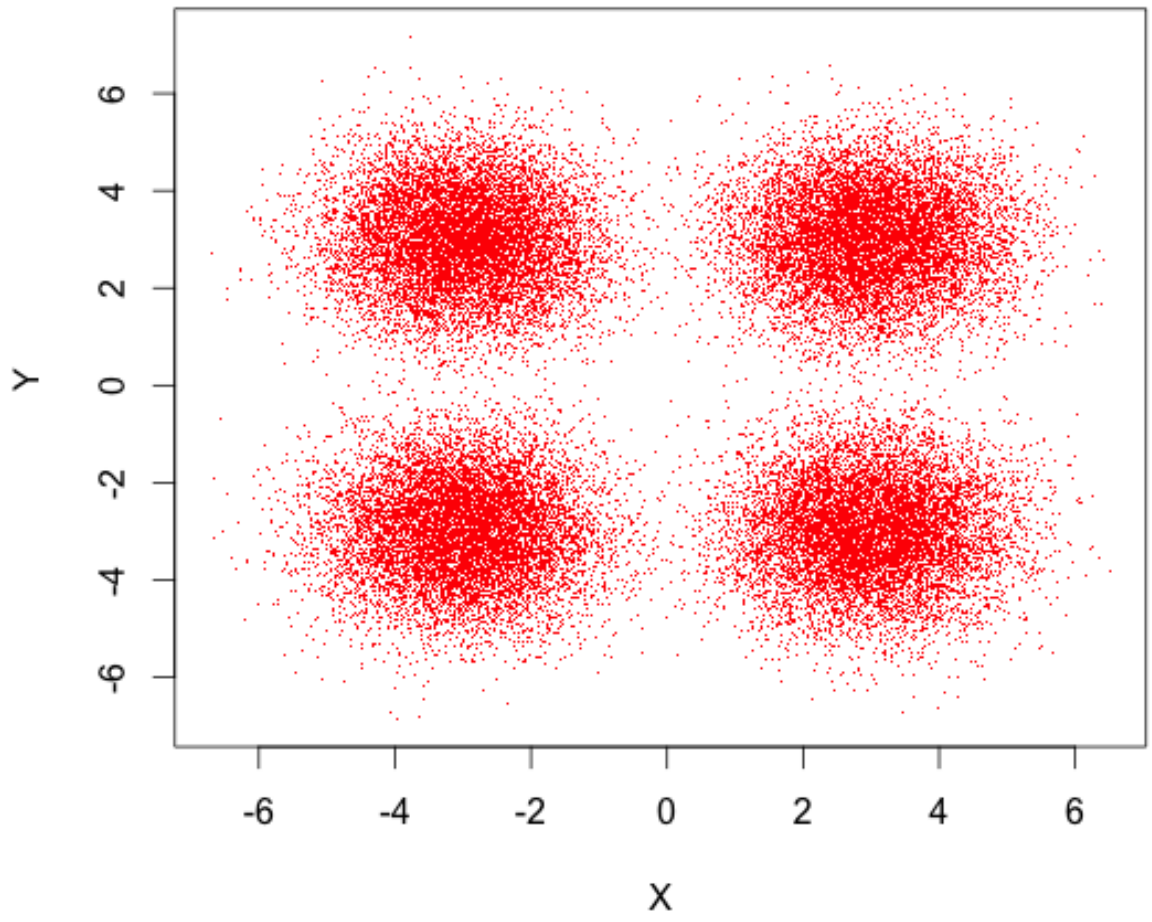
$$\left.\begin{bmatrix} x_{1,1} \\ \vdots \\ x_{1,p} \end{bmatrix} \quad \begin{bmatrix} x_{n,1} \\ \vdots \\ x_{n,p} \end{bmatrix}\right\} p \text{ coordinates for each observation : } x_i \in \mathbb{R}^p$$

**Examples** : $k$-means, EM, hierarchical clustering

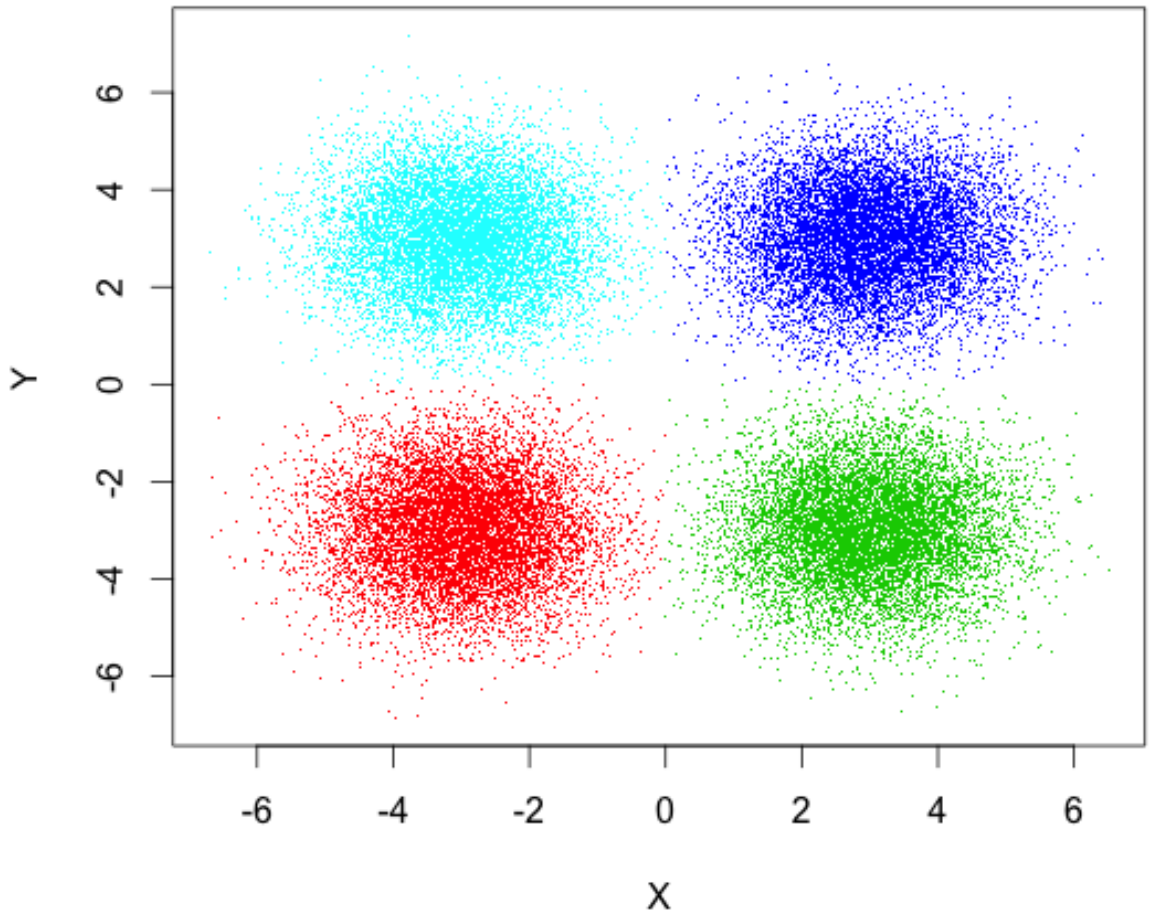
# Example : $k$-means



# Example : $k$-means



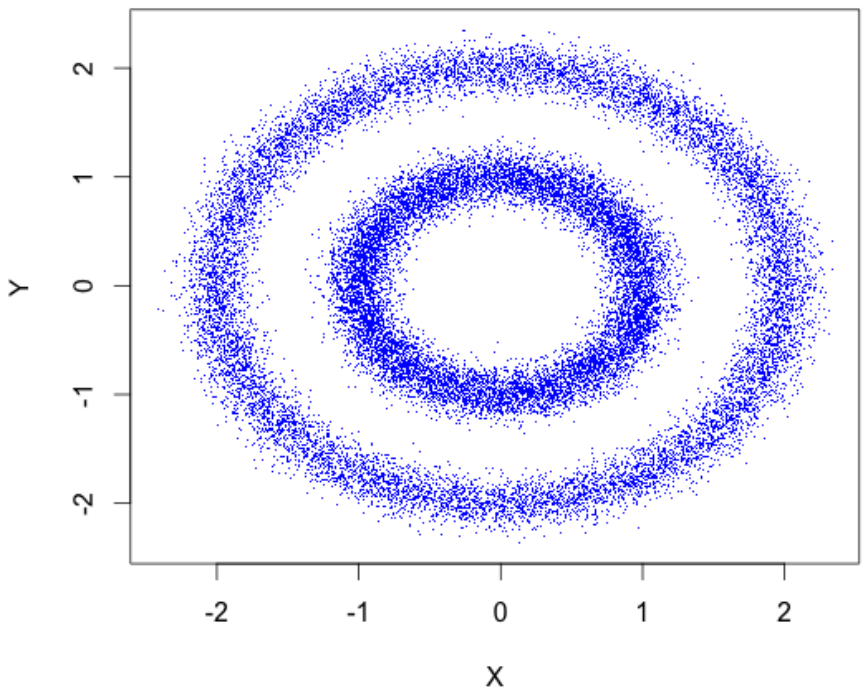
# Example : $k$-means

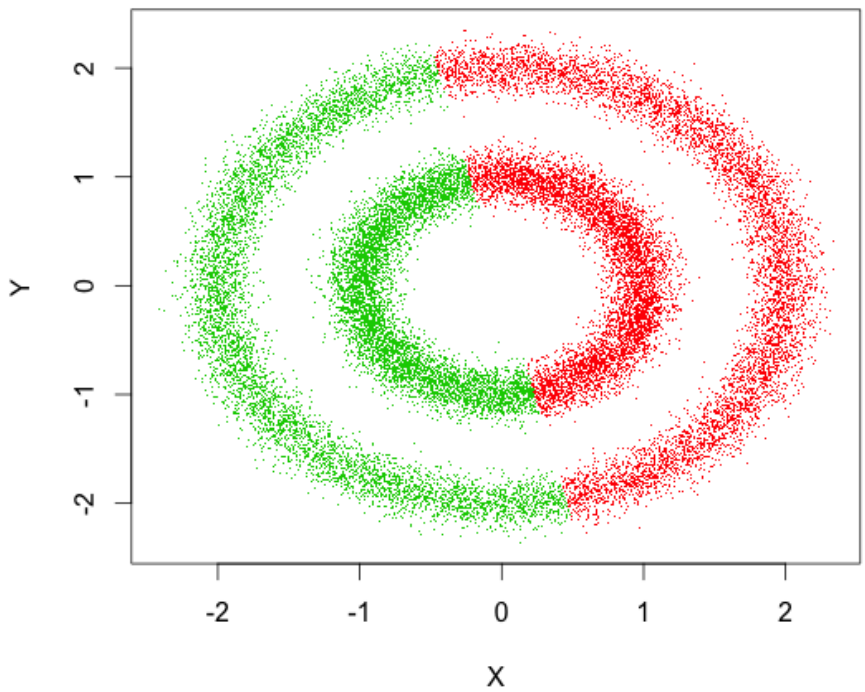


# Example : $k$-means

Disadvantages :

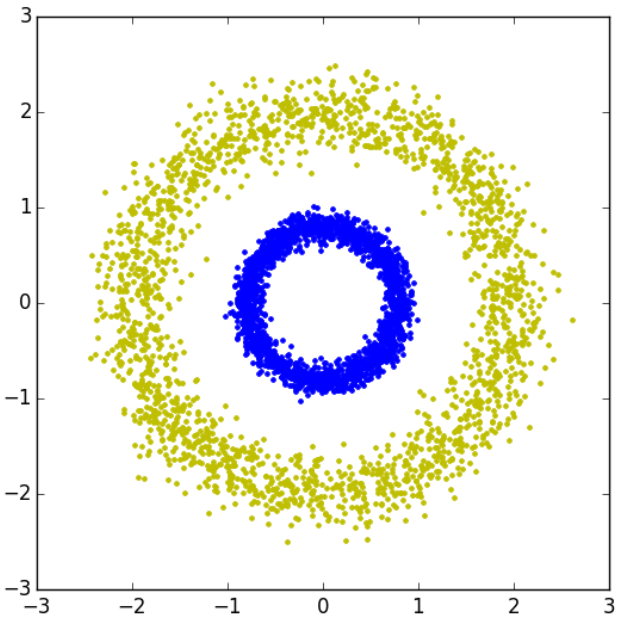
- Not efficient in large dimension (when $p \gg 1$)
- Even in low dimension : means are not always relevant :

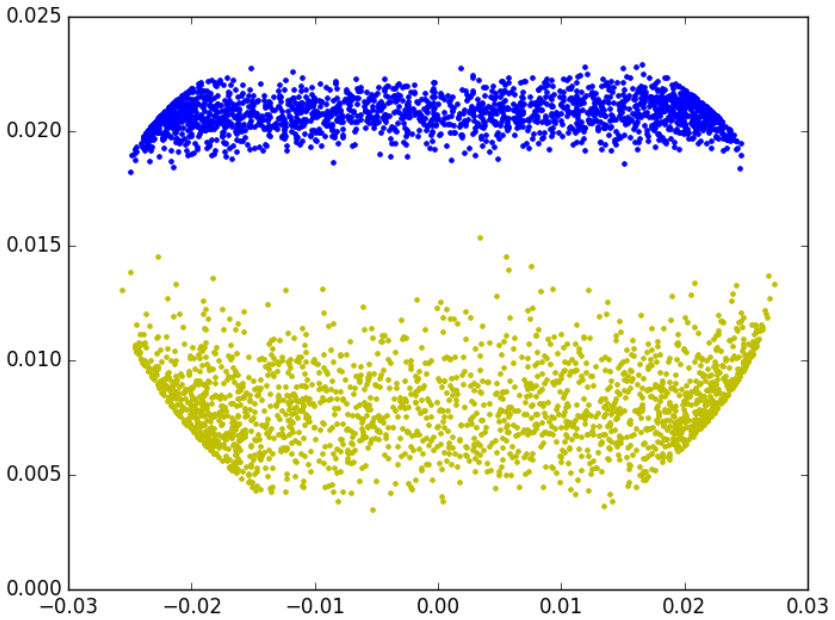


# Spectral clustering : example and principle (1)

First **transform** the observations :



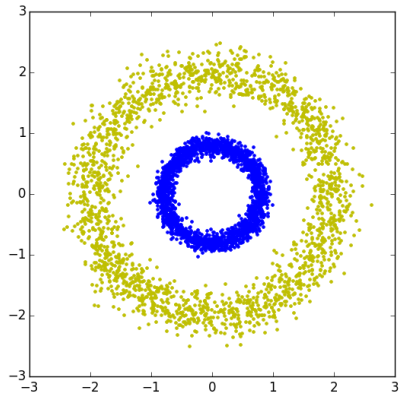




Left : $x_1, \ldots, x_n \implies K := \left[e^{-c\|x_j - x_i\|^2}\right]_{i,j=1}^{n}$ with **eigenvectors**

$\vec{V}_1 \geq \vec{V}_2 \geq \cdots \implies$ right : $y_1, \ldots, y_n$ defined by :

$$\begin{bmatrix}\vec{V}_2 & \vec{V}_1\end{bmatrix} =: \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

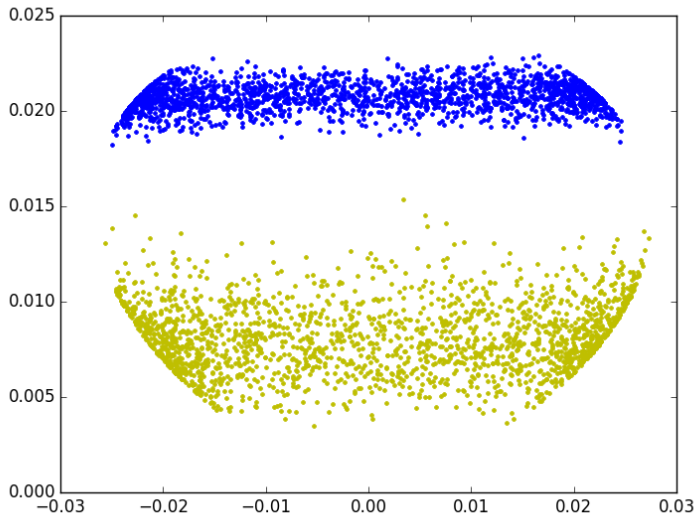
# Spectral clustering : example and principle (1)



- $x_1, \ldots, x_{n_1}$ : small circle
- $x_{n_1+1}, \ldots, x_{n_1+n_2}$ : large circle

$\hookrightarrow$ Matrix $K := \left[e^{-c\|x_j - x_i\|^2}\right]_{i,j=1}^{n}$ : $\quad K = \begin{bmatrix} X & \varepsilon \\ \varepsilon & Y \end{bmatrix} \approx \begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix}$

$\hookrightarrow$ Maximal eigenvectors : $\approx$ either supported by $\vec{e}_1, \ldots, \vec{e}_{n_1}$ or by $\vec{e}_{n_1+1}, \ldots, \vec{e}_{n_1+n_2}$



# Spectral clustering : principle (2)

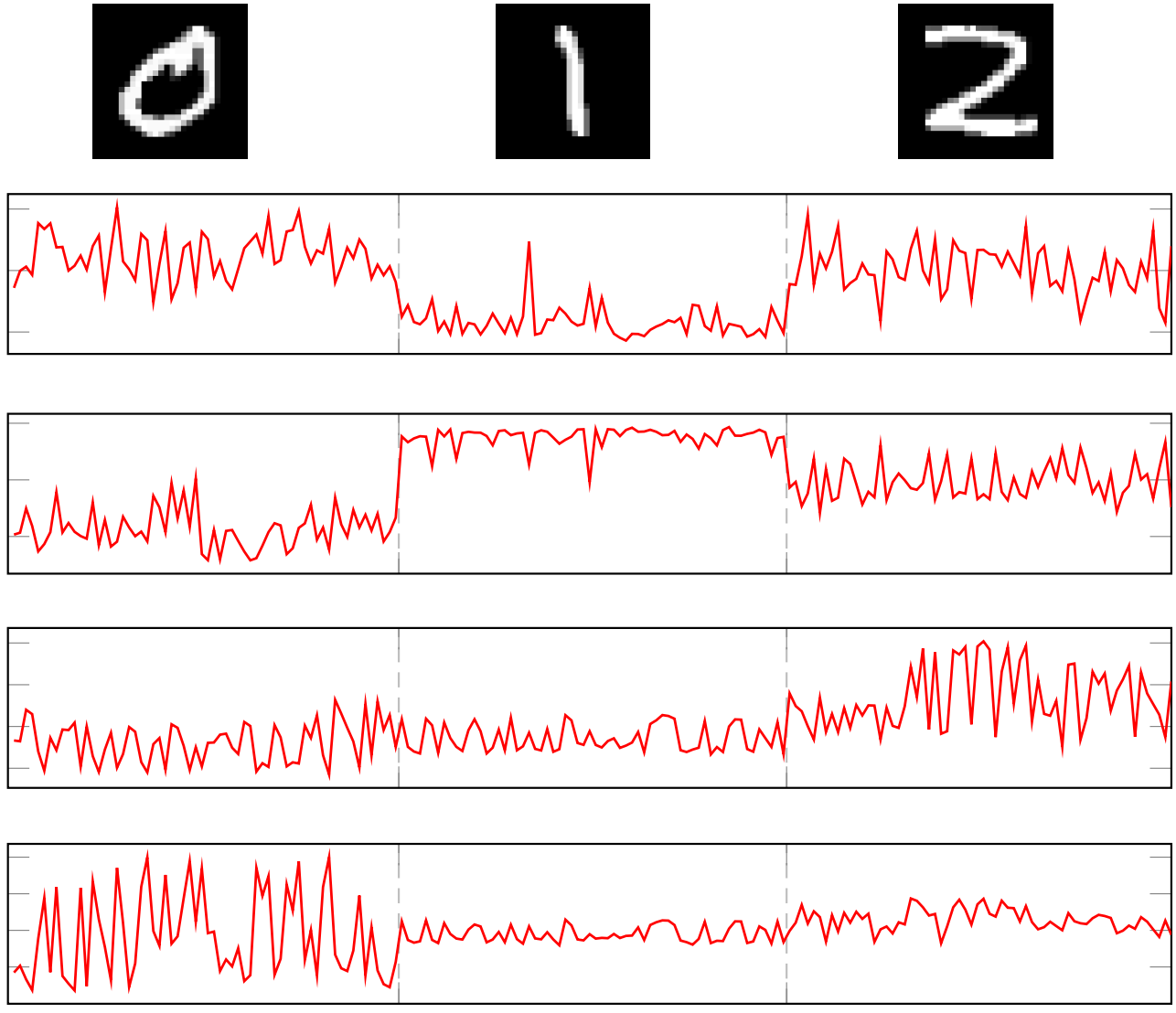
$f(\|x_j - x_i\|) \geq 0$ : **similarity** of $x_i$ and $x_j$

(ex : $f(\|x_j - x_i\|) = \mathrm{e}^{-c\|x_j - x_i\|^2}$)

$$L = \left[\frac{f(\|x_j - x_i\|)}{\sqrt{d_i d_j}}\right]_{i,j=1}^{n}, \quad d_i := \sum_k f(\|x_k - x_i\|)$$

## Spectral clustering of $x_1, \ldots, x_n$ in $k$ classes (2) :

*$L$ : symmetric Laplacian matrix. Replace observations $x_1, \ldots, x_n$ by the largest eigenvectors of $L$ and apply $k$-means on these (new) observations.*



: Four leading eigenvectors of $L$ for (partial) MNIST data ($n = 192$, $p = 784$, $k = 3$)

# Goal and method

**Goal :** develop mathematical analysis of the algorithm for $n, p \gg 1$ :

- Phase transitions (when is the clustering working ?)
- Content of each eigenvector ?
- Influence of the kernel function $f$
- Assess algorithm performance

**Method :** use statistical hypotheses (Gaussian mixture et independence of the observations) and recent Random Matrix Theory technics (*spiked models, BBP phase transition*)

# Model and Assumptions

**Gaussian mixture model :**

- $x_1, \dots, x_n \in \mathbb{R}^p$ independent
- $k$ classes $\mathcal{C}_1, \dots, \mathcal{C}_k$
- $\mathcal{C}_a = \{x \mid x \sim \mathcal{N}(\mu_a, C_a)\}, \quad \|C_a\| = O(1)$

**Convergence rate :** As $n \to \infty$,

1. **Data scaling** : $c_0 := \frac{p}{n}$ away from $0$ and $+\infty$
2. **Class scaling** : $c_a := \frac{\#\mathcal{C}_a}{n}$ away from $0$ and $1$
3. **Mean and covariance scaling** :
   Cases where **simple methods** are efficient :
   - $p \gg 1 \implies x_i - \mu_a = O(\sqrt{\operatorname{Tr} C_a}) = O(\sqrt{p})$ so
     $\|\mu_a - \mu_b\| \gg \sqrt{p} \implies k$-means (possibly *well* projected) is efficient
   - $\|\mu_a\| \ll \sqrt{p}$ et $|\operatorname{Tr} C_a - \operatorname{Tr} C_b| \gg \sqrt{p} \implies k$-means on $\|x_i\|$ is efficient

# Model and Assumptions

**Gaussian mixture model :**

- $x_1, \dots, x_n \in \mathbb{R}^p$ independent
- $k$ classes $\mathcal{C}_1, \dots, \mathcal{C}_k$
- $\mathcal{C}_a = \{x \mid x \sim \mathcal{N}(\mu_a, C_a)\}, \quad \|C_a\| = O(1)$

**Convergence rate :** As $n \to \infty$,

1. **Data scaling** : $c_0 := \frac{p}{n}$ away from $0$ and $+\infty$
2. **Class scaling** : $c_a := \frac{\#\mathcal{C}_a}{n}$ away from $0$ and $1$
3. **Mean scaling** : with $\sum_{a=1}^k c_a \mu_a = \vec{0}$, $\|\mu_a\| = O(1)$
4. **Covariance scaling** : with $C^\circ := \sum_{a=1}^k c_a C_a$ and $C_a^\circ := C_a - C^\circ$, we have

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

Then $\frac{1}{p}\|x_i - x_j\|^2 \approx \tau := \frac{2}{p} \operatorname{Tr} C^\circ$

# Matrix of interest

- Kernel matrix :

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

for some sufficiently smooth nonnegative $f$

- We study the normalized **Laplacian** matrix :

$$L = nD^{-\frac{1}{2}} K D^{-\frac{1}{2}}$$

with $D = \mathrm{diag}(d_i, 1 \leq i \leq n)$, $d_i = \sum_j K_{ij}$.

# Objectives

We want to derive, for each leading eigenvector $\vec{V}$ and each class $\mathcal{C}_a$, $a = 1, \ldots, k$ :

- *Class-wise eigenvector means* :

$$\alpha_a(\vec{V}) := \frac{1}{\#\mathcal{C}_a} \left\langle \vec{V}, 1_{\mathcal{C}_a} \right\rangle$$

- *Class-wise eigenvector fluctuations* :

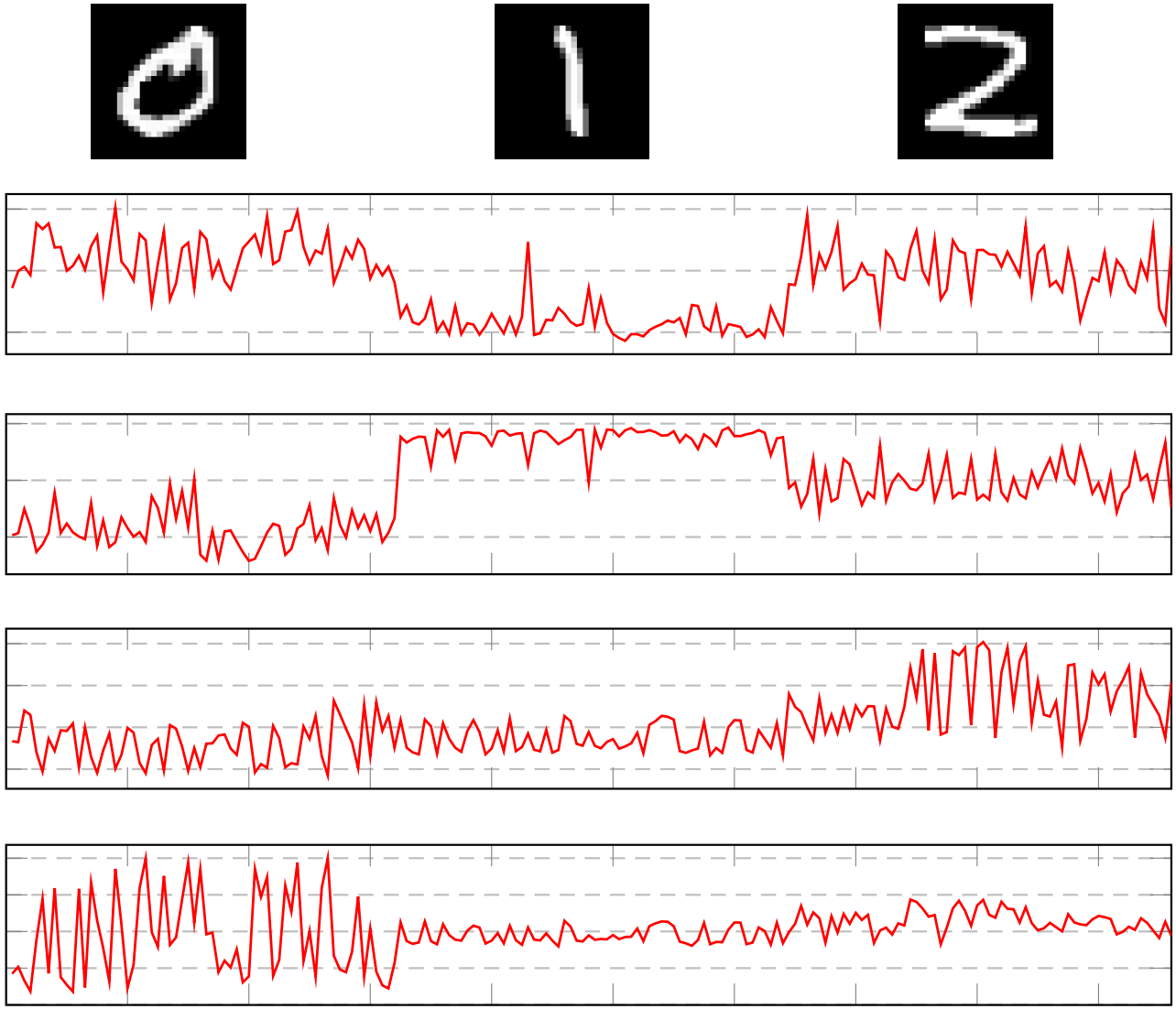
$$\left\| \mathrm{diag}(1_{\mathcal{C}_a}) \left( \vec{V} - \alpha_a(\vec{V}) 1_{\mathcal{C}_a} \right) \right\|$$

- *Class-wise cross correlations* :

$$\left\langle \left( \vec{V} - \alpha_a(\vec{V}) 1_{\mathcal{C}_a} \right), \mathrm{diag}(1_{\mathcal{C}_a}) \left( \vec{W} - \alpha_a(\vec{W}) 1_{\mathcal{C}_a} \right) \right\rangle$$

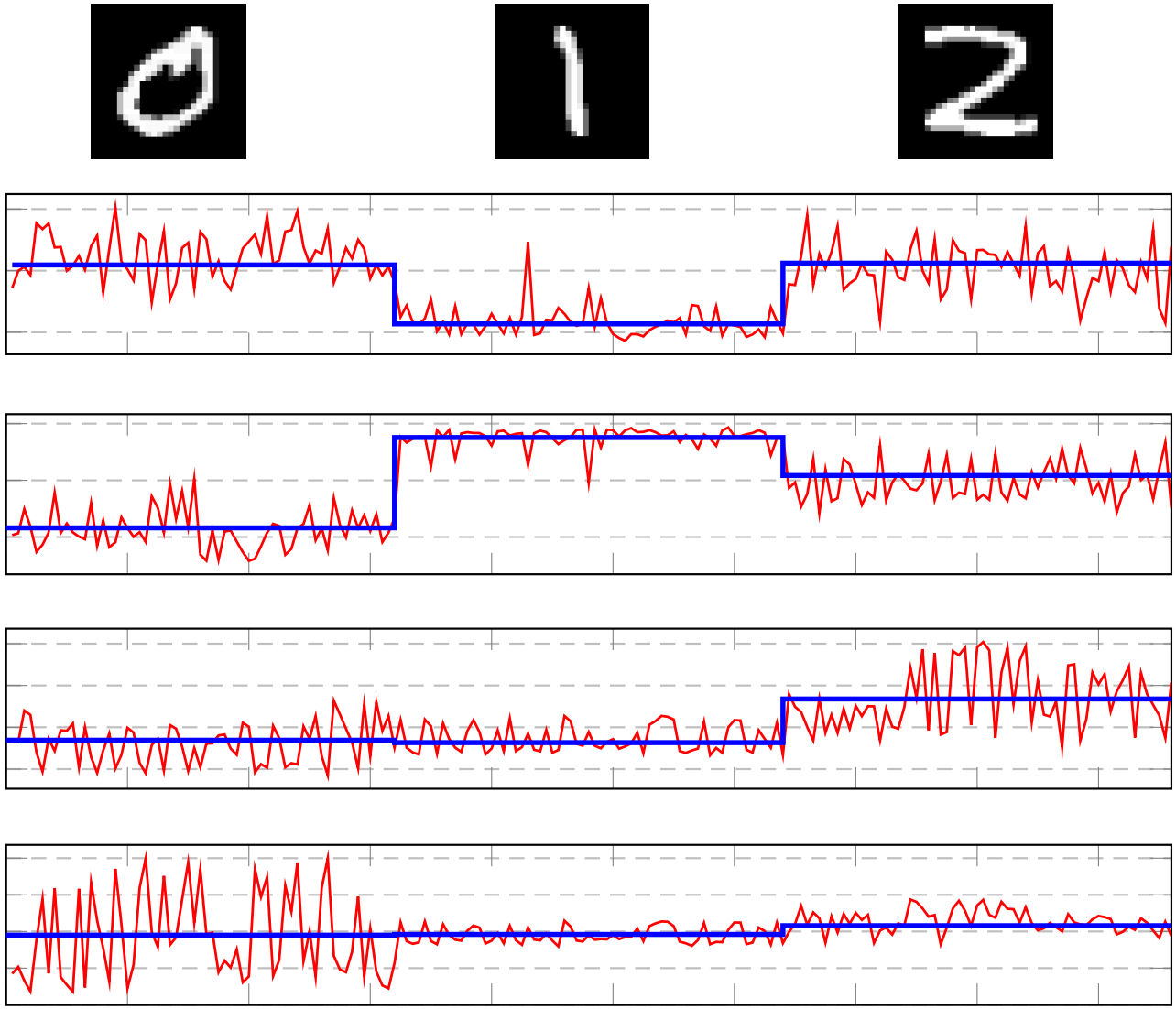
  for $\vec{W}$ another leading eigenvector

# Class-wise eigenvector means and fluctuations



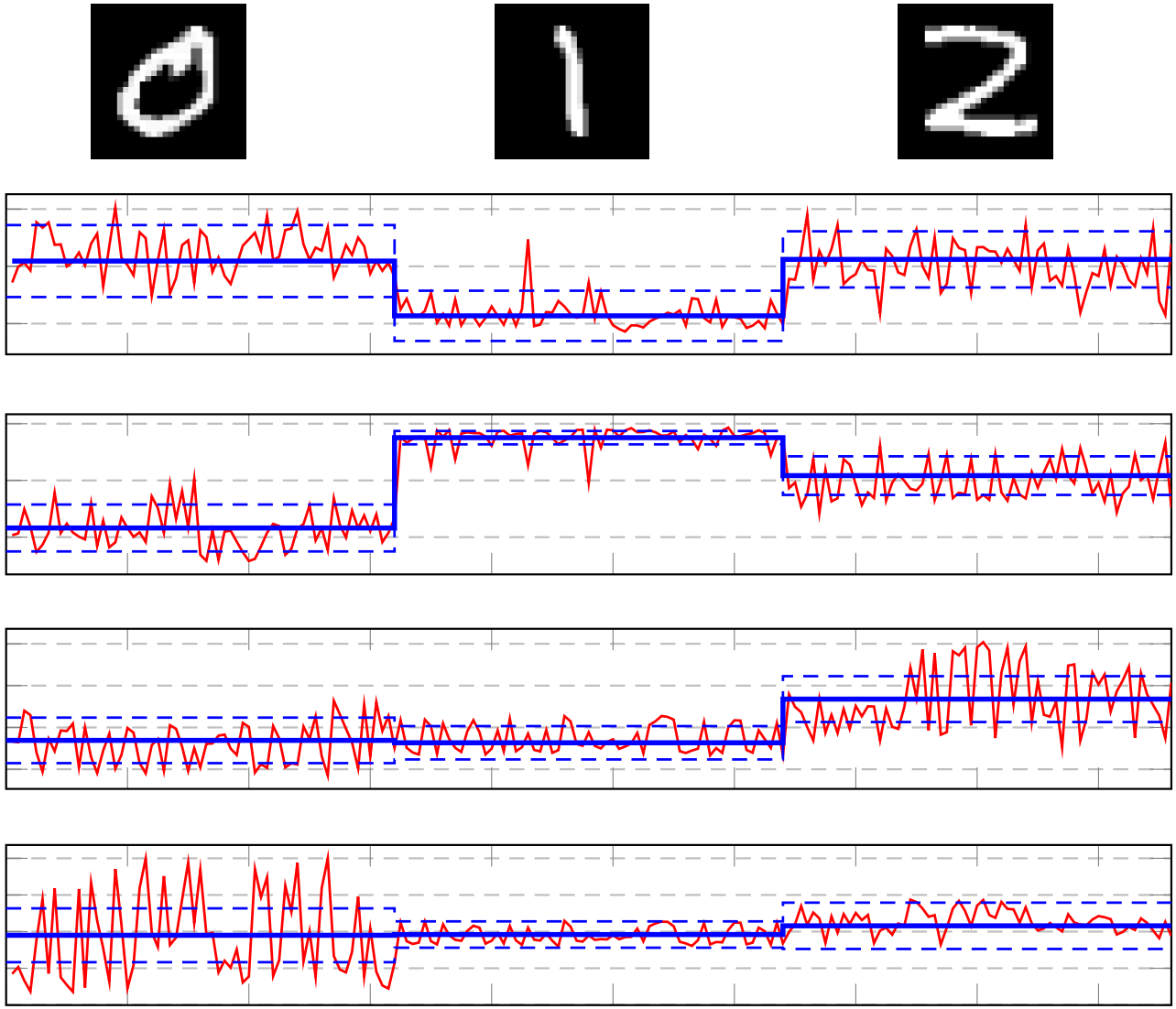
: MNIST data : four leading eigenvectors of $L$ (**red**), versus $\hat{L}$ (**black**) and theoretical findings (**blue**).

# Class-wise eigenvector means and fluctuations



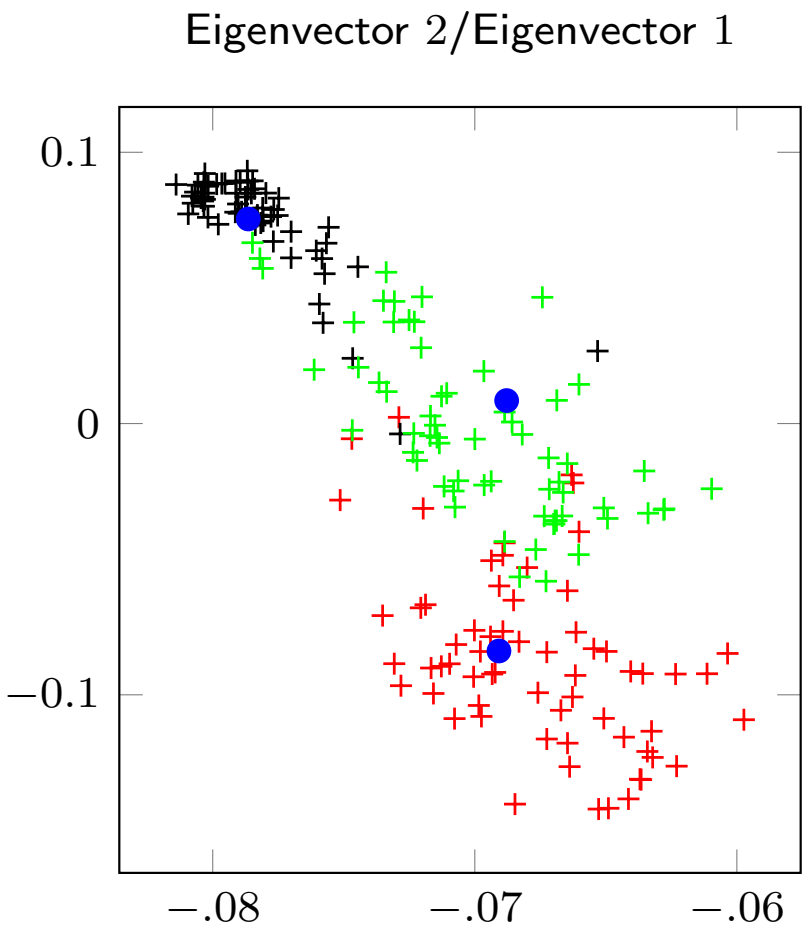
: MNIST data : four leading eigenvectors of $L$ (**red**), versus $\hat{L}$ (**black**) and theoretical findings (**blue**).

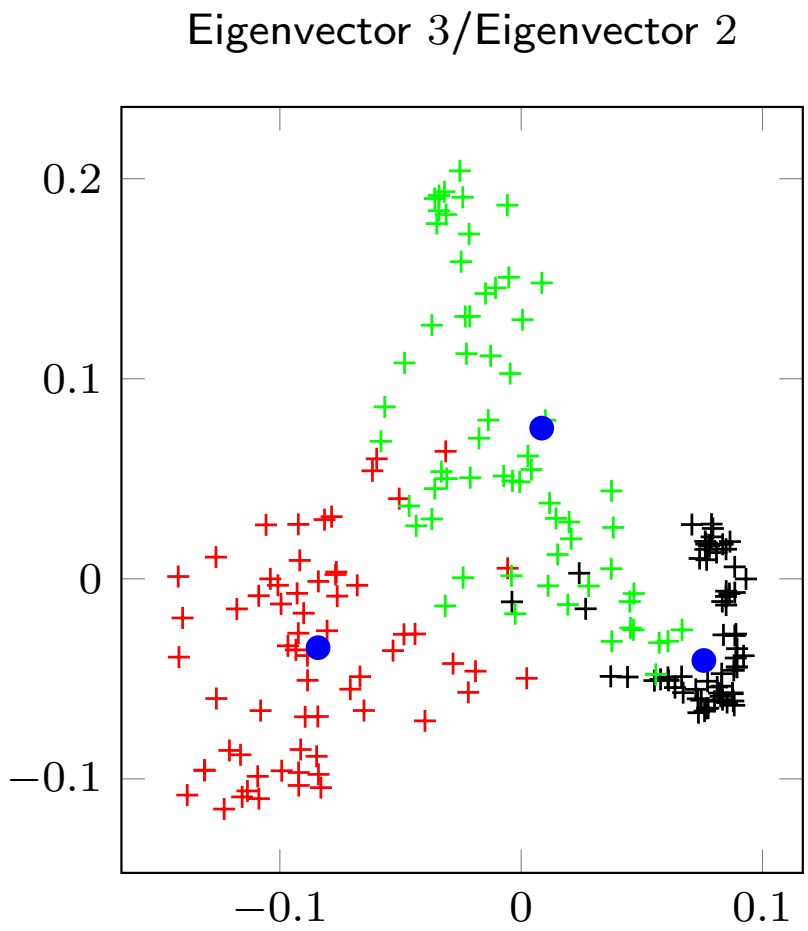
# Class-wise eigenvector means and fluctuations



: MNIST data : four leading eigenvectors of $L$ (**red**), versus $\hat{L}$ (**black**) and theoretical findings (**blue**).

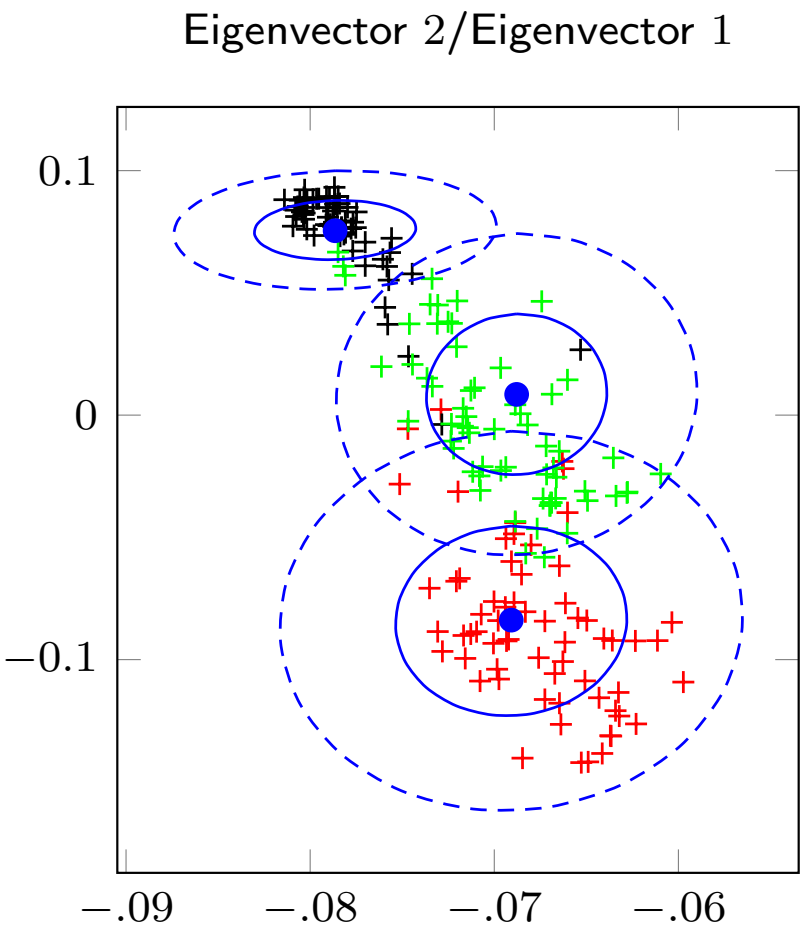
# Class-wise means, fluctuations and cross correlations

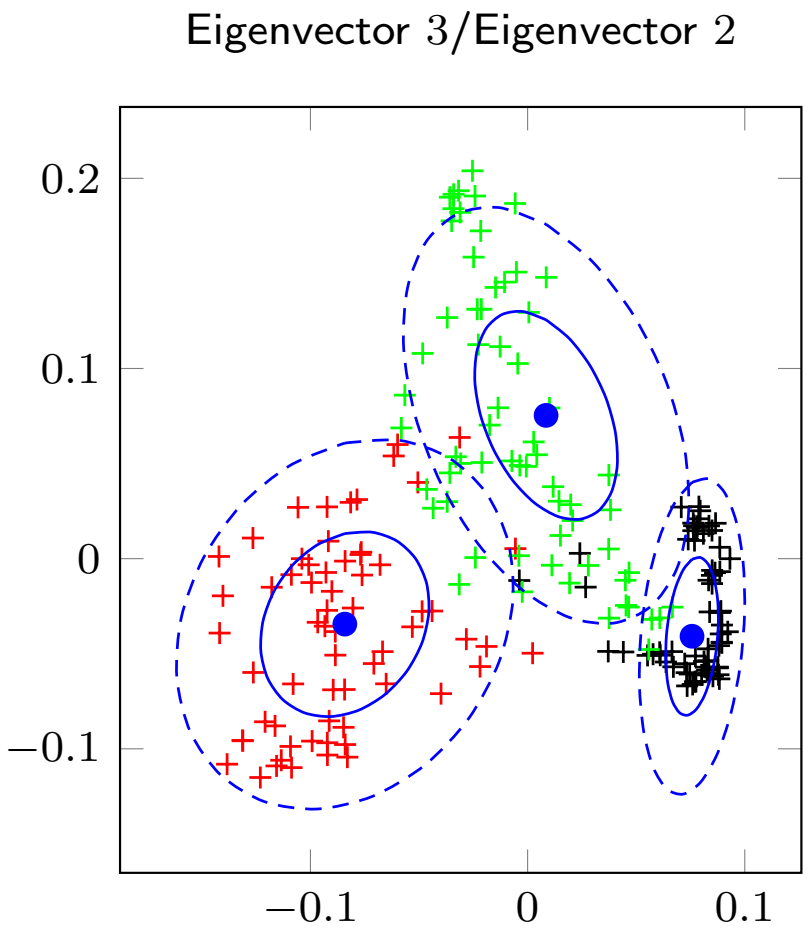



: 2D representation of eigenvectors of $L$, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

# Class-wise means, fluctuations and cross correlations




: 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**.

# Matrix of interest

**Observations** : Spectrum of $L = nD^{-\frac{1}{2}}KD^{-\frac{1}{2}}$ :

a) Dominant eigenvalue $n$ with eigenvector $(\sqrt{d_1}, \ldots, \sqrt{d_n})^\mathsf{T} = D^{\frac{1}{2}}1_n$ (with $1_n = (1, \ldots, 1)^\mathsf{T}$)

b) All other eigenvalues of order $O(1)$, with some isolated eigenvalues with eigenvectors containing information about the classes

$\Rightarrow$ Naturally leads to study :

- Dominant eigenvector :

$$\frac{D^{\frac{1}{2}}1_n}{\sqrt{1_n^\mathsf{T} D 1_n}} = \frac{(\sqrt{d_1}, \ldots, \sqrt{d_n})^\mathsf{T}}{\sqrt{d_1 + \cdots + d_n}}$$

- Projected normalized Laplacian :

$$L' = nD^{-\frac{1}{2}}KD^{-\frac{1}{2}} - n\frac{D^{\frac{1}{2}}1_n 1_n^\mathsf{T} D^{\frac{1}{2}}}{1_n^\mathsf{T} D 1_n}.$$

# Eigenvectors

**Dominant Eigenvector** :

## Proposition (Eigenvector $D^{\frac{1}{2}}1_n$)

*We have*

$$\frac{D^{\frac{1}{2}}1_n}{\sqrt{1_n^\intercal D 1_n}} = \frac{1_n}{\sqrt{n}} + \frac{1}{n\sqrt{c_0}}\frac{f'(\tau)}{2f(\tau)}\left[\{t_a 1_{\mathcal{C}_a}\}_{a=1}^k + \operatorname{diag}\left\{\sqrt{\frac{2}{p}\operatorname{Tr}(C_a^2)}1_{\mathcal{C}_a}\right\}_{a=1}^k \varphi\right] + o(n^{-1})$$

*with* $t_a := \frac{1}{\sqrt{p}}\operatorname{Tr} C_a^\circ$ $(a = 1, \ldots, k)$ *and* $\varphi \sim \mathcal{N}(0, I_n)$.

**Remark :**

- structure of $D^{\frac{1}{2}}1_n$ : block-wise constant + noise
- only information about $\operatorname{Tr} C_a^\circ$ !

# Random Matrix Equivalent

## Theorem (Random Matrix Equivalent)

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

$$\hat{L}' = PW^{\mathsf{T}}WP + \chi,$$

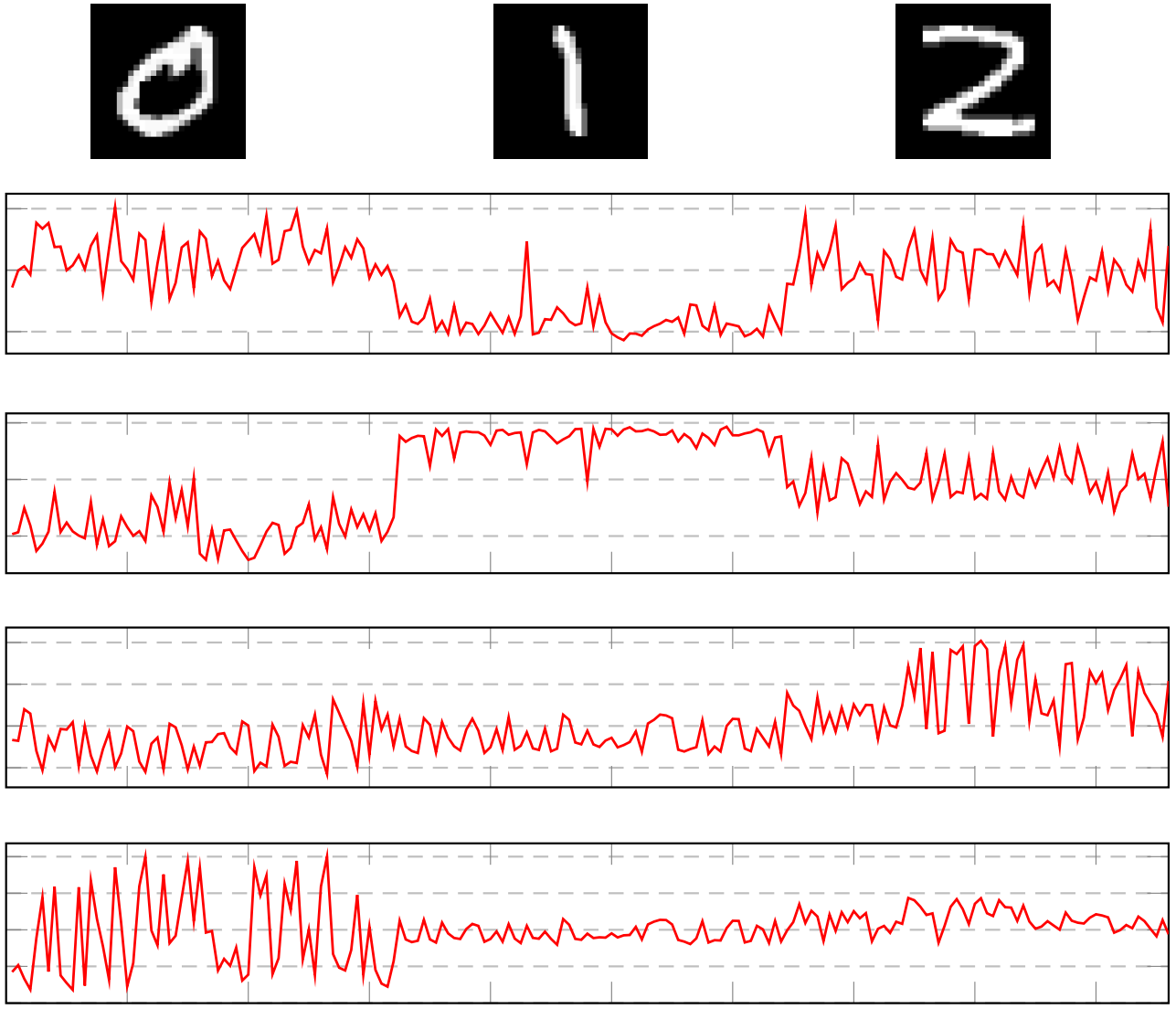
*with* $P$ *orthogonal projection onto* $\{x_1 + \cdots + x_n = 0\}$,
$W = [w_1, \ldots, w_n] \in \mathbb{R}^{p \times n}$ *Gaussian* ($x_i = \mu_a + p^{1/2} w_i$) *and*

$$\chi \text{ is matrix with rank} \leq 2k + 4$$

(***spiked model***) *depending on :*

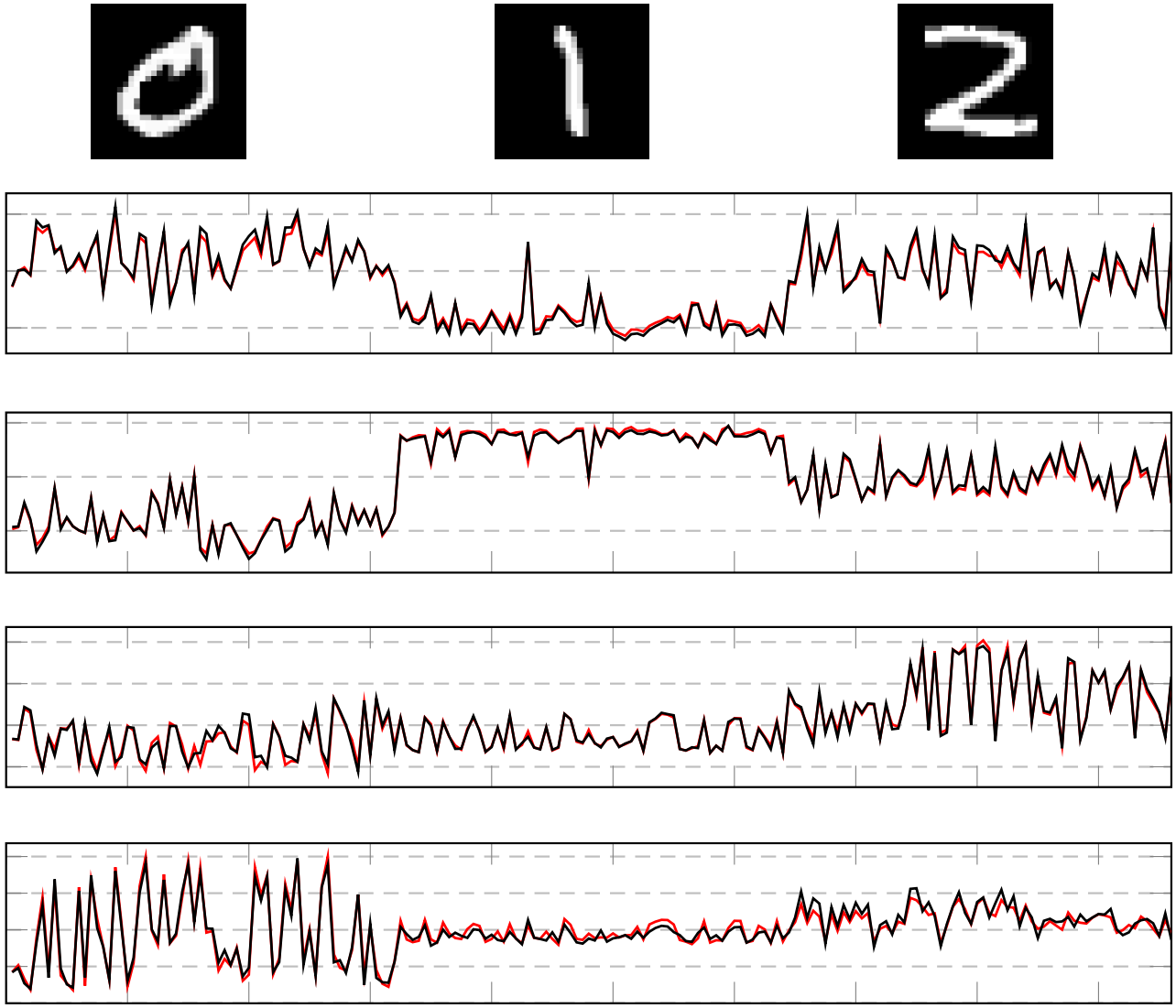
- *the class structure*
- *the function* $f$ *through the numbers* $f(\tau)$, $f'(\tau)$, $f''(\tau)$, $\tau = \frac{2}{p} \operatorname{Tr} C^\circ$
- *the means* $\mu_a$ $(a = 1, \ldots, k)$
- *the traces* $t_a = \frac{1}{\sqrt{p}} \operatorname{Tr} C_a^\circ$ $(a = 1, \ldots, k)$
- *the cross-traces* $T_{a,b} := \frac{1}{p} \operatorname{Tr} C_a^\circ C_b^\circ$ $(a, b = 1, \ldots, k)$

# Equivalence between $L$ and $\hat{L}$ : eigenvectors



: MNIST data : four leading eigenvectors of $L$ (**red**), versus $\hat{L}$ (**black**)

# Equivalence between $L$ and $\hat{L}$ : eigenvectors



: MNIST data : four leading eigenvectors of $L$ (**red**), versus $\hat{L}$ (**black**)

# Towards the eigenvectors of $\hat{L}' = PW^{\mathsf{T}}WP + \chi$

1. Study eigenvalue distribution (and its limit support $\mathcal{S}$) of $PW^{\mathsf{T}}WP$

2. Isolated eigenvalues of $\hat{L}'$ : solve, for $z \notin \mathcal{S}$,

$$\det\left(PW^{\mathsf{T}}WP + \chi - z\right) \;=\; 0$$

turning this $n \times n$ determinant to a smaller one :

$$\det\left(PW^{\mathsf{T}}WP + \chi - zI_n\right)$$

$$= \det\left(PW^{\mathsf{T}}WP - z\right) \det \underbrace{\left(1 + \left(PW^{\mathsf{T}}WP - z\right)^{-1}\chi\right)}_{\text{matrix with small co-rank}}$$

3. Study the eigenvectors thanks to the Cauchy Formula :

$$\text{Spectral Projection of } \hat{L}' \text{ on } I \;=\; \frac{1}{2\mathrm{i}\pi}\oint_{\gamma_I}(z - \hat{L})^{-1}\mathrm{d}z$$

# Step 1 : eigenvalue distribution of $PW^\mathsf{T}WP$

• Random measure $\mu_n = n^{-1}\sum_{i=1}^n \delta_{\lambda_i(PW^\mathsf{T}WP)}$ studied through its **Stieltjes transform**

$$S_{\mu_n}(z) \;=\; \int_{\lambda\in\mathbb{R}} \frac{\mu_n(\mathrm{d}\lambda)}{\lambda - z} \;=\; \frac{1}{n}\operatorname{Tr}(Q_z) \qquad \text{for } Q_z = (PW^\mathsf{T}WP - z)^{-1}$$

• $Q_z = z^{-1}PW^\mathsf{T}WPQ_z - z^{-1}I_n$ + Stein Formula for Gaussian variables $\mathbb{E}[Xf(X)] = \sigma^2\mathbb{E}[f'(X)] \Longrightarrow$ Loop equations for $Q_z \Longrightarrow$ fixed point characterization of a deterministic equivalent :

$$Q_z \;=\; \frac{p}{n}\operatorname{diag}\left\{g_a(z)1_{n_a}\right\}_{a=1}^k + o(1)$$

for $(g_a(z))_{a=1,\ldots,k} \in \mathbb{C}^k$ solution of

$$g_a(z) \;=\; \frac{1}{\frac{1}{n}\operatorname{tr} C_a\left(I_p + \sum_{b=1}^k \frac{n_b}{n} g_b(z) C_b\right)^{-1} - pz/n} \qquad (a = 1, \ldots, k)$$

$\Longrightarrow$ deterministic equivalent of $\mu_n$

# Step 2 : isolated eigenvalues of $\hat{L}'$

$$\hat{L}' \quad = \quad \underset{\substack{\downarrow \\ \text{well understood}}}{PW^{\mathsf{T}}WP} \quad + \quad \underset{\substack{\downarrow \\ \text{bounded rank}}}{\chi}$$

$\hookrightarrow$ classical framework of **spiked random matrices theory**

## Theorem

*There is a (complicated but human) function $F(z)$ such that (up to some technical hypotheses) the isolated eigenvalues of $\hat{L}'$ are the roots of*

$$F(z) = 0.$$

# Step 3 : isolated eigenvectors of $\hat{L}'$

We want to derive, for each leading eigenvector $\vec{V}$ and each $a = 1, \dots, k$ :

- *Class-wise eigenvector means* :

$$\alpha_a(\vec{V}) := \frac{1}{\#\mathcal{C}_a} \left\langle \vec{V}, 1_{\mathcal{C}_a} \right\rangle$$

- *Class-wise eigenvector fluctuations* :

$$\left\| \mathrm{diag}(1_{\mathcal{C}_a}) \left( \vec{V} - \alpha_a(\vec{V}) 1_{\mathcal{C}_a} \right) \right\|$$

- *Class-wise cross correlations* :

$$\left\langle \left( \vec{V} - \alpha_a(\vec{V}) 1_{\mathcal{C}_a} \right), \mathrm{diag}(1_{\mathcal{C}_a}) \left( \vec{W} - \alpha_a(\vec{W}) 1_{\mathcal{C}_a} \right) \right\rangle$$

  for $\vec{W}$ another leading eigenvector

$\hookrightarrow$ one needs, for $\Pi = \vec{V}\vec{V}^{\mathsf{T}}$ and $\Pi' = \vec{W}\vec{W}^{\mathsf{T}}$, the numbers

$$p^{-1} J^{\mathsf{T}} \Pi J \quad ; \quad p^{-1} J^{\mathsf{T}} \Pi \, \mathrm{diag}(1_{\mathcal{C}_a}) \Pi' J \qquad (1 \leq a \leq k)$$

for $J = [1_{\mathcal{C}_1} \cdots 1_{\mathcal{C}_k}] \in \mathbb{R}^{n \times k}$.

# Step 3 : isolated eigenvectors of $\hat{L}'$

- Cauchy Formula :

$$\text{Spectral Projection of } \hat{L}' \text{ on } I \;=\; -\frac{1}{2\mathrm{i}\pi}\oint_{\gamma_I}(\hat{L}'-z)^{-1}\mathrm{d}z$$

- Woodbury matrix identity : for $Q_z=(PW^{\mathsf{T}}WP-z)^{-1}$,

$$(\hat{L}'-z)^{-1} \;=\; Q_z - Q_zU(B^{-1}+U^{\mathsf{T}}Q_zU)^{-1}U^{\mathsf{T}}Q_z$$

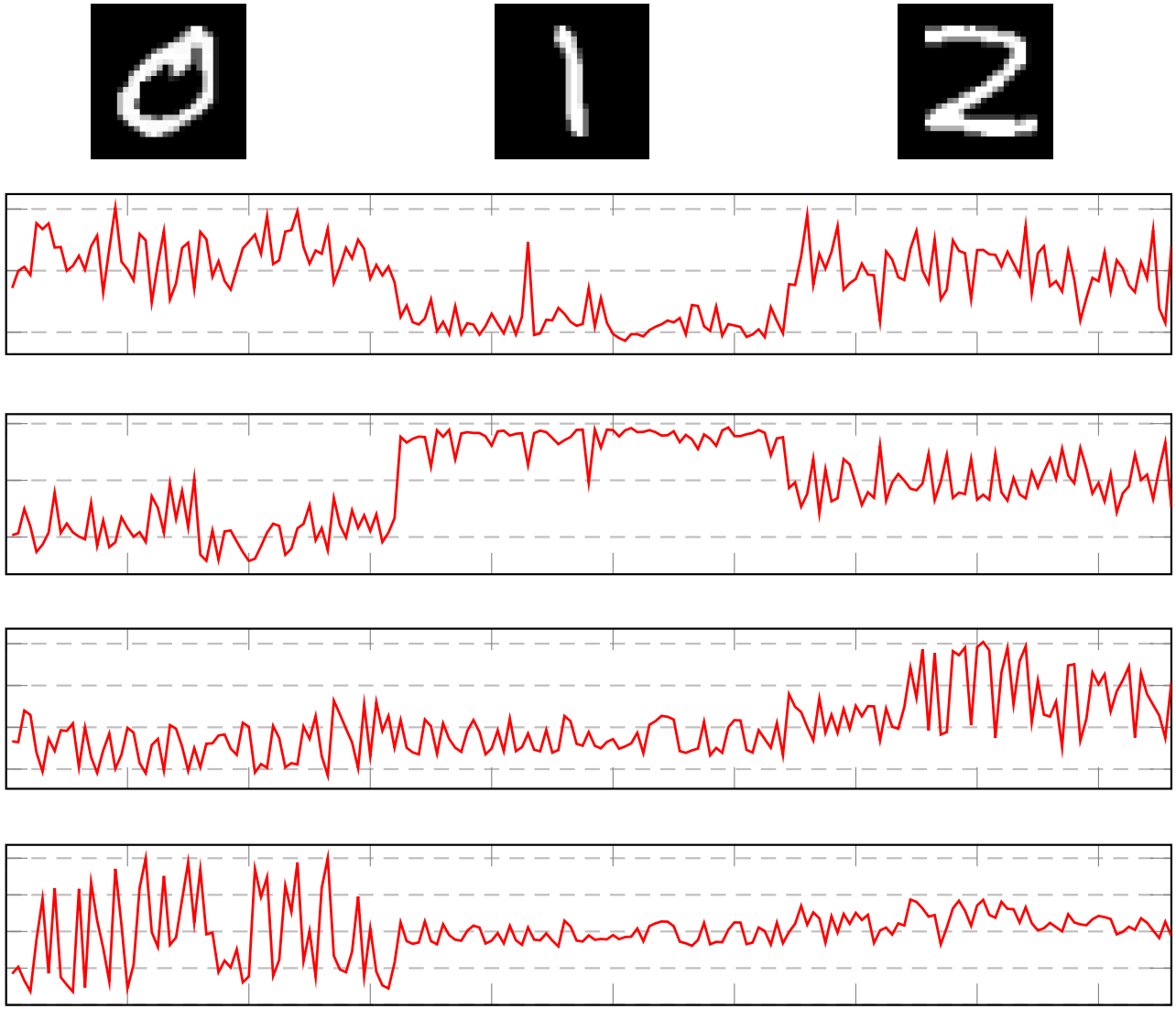
## Theorem

*The limits of the numbers*

$$p^{-1}J^{\mathsf{T}}\Pi J \qquad ; \qquad p^{-1}J^{\mathsf{T}}\Pi\,\mathrm{diag}(1_{\mathcal{C}_a})\Pi' J \qquad (1\le a\le k)$$

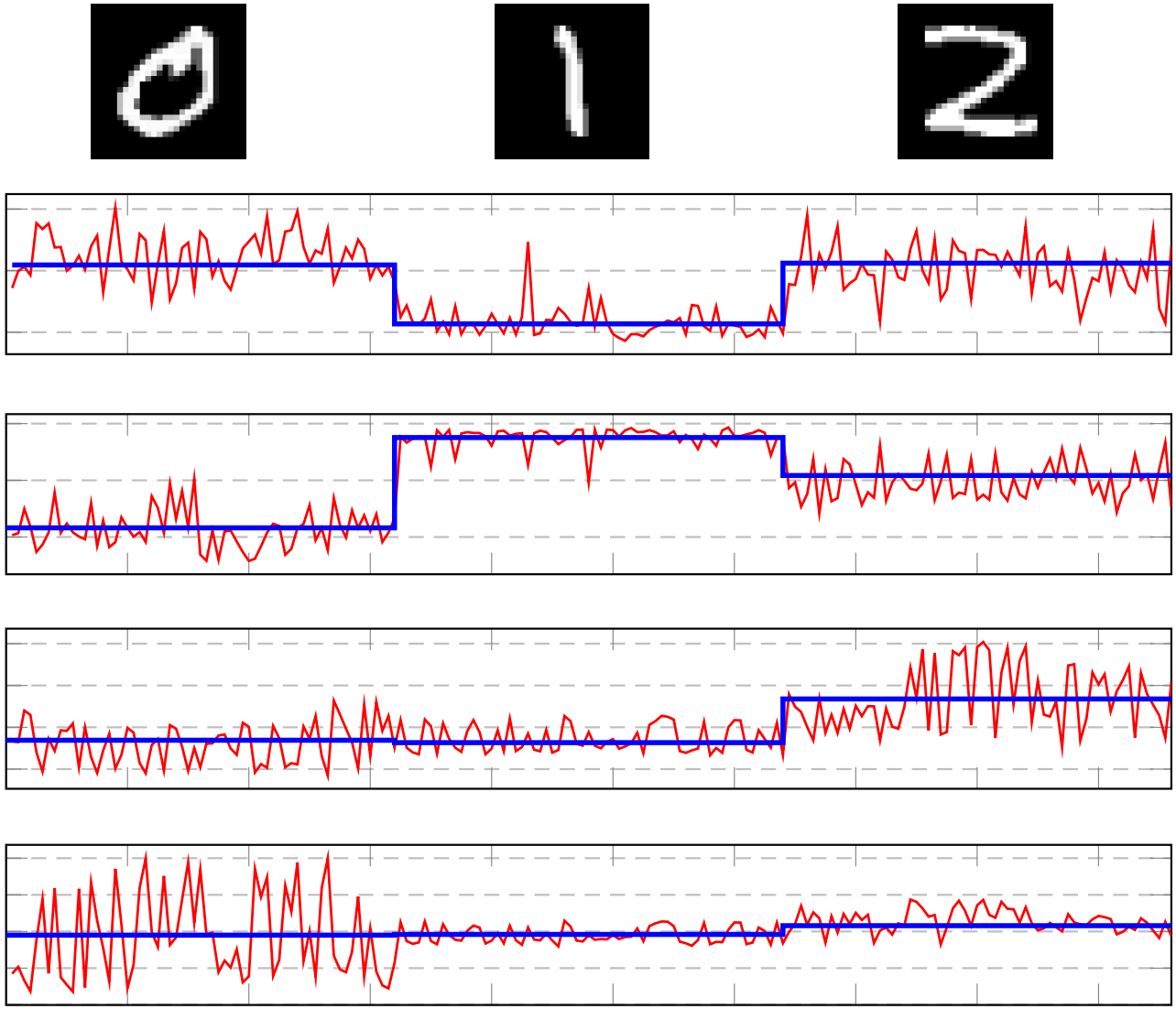
*can be computed.*

# Class-wise eigenvector means and fluctuations



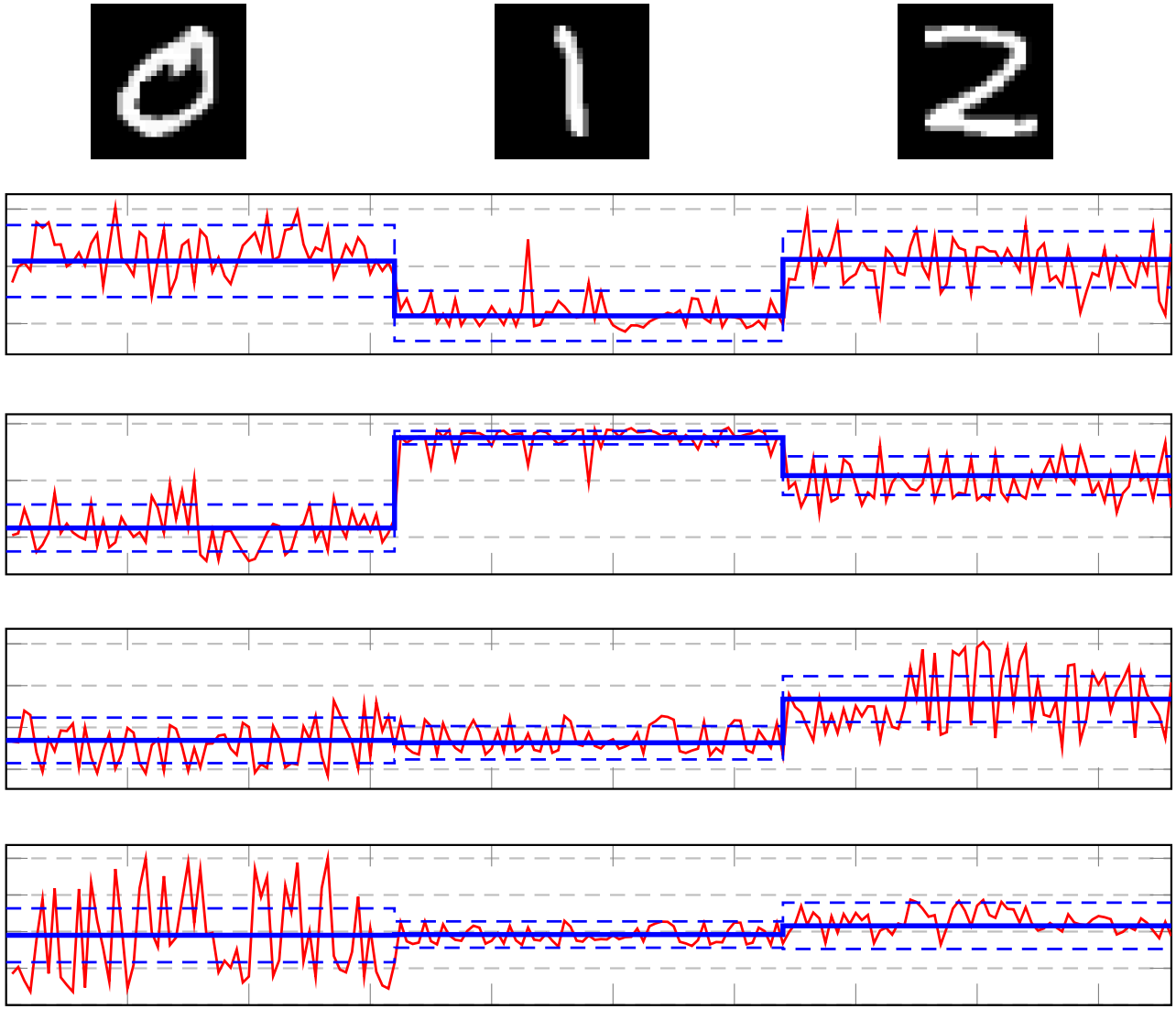
: MNIST data : four leading eigenvectors of $L$ (**red**), versus $\hat{L}$ (**black**) and theoretical findings (**blue**).

# Class-wise eigenvector means and fluctuations



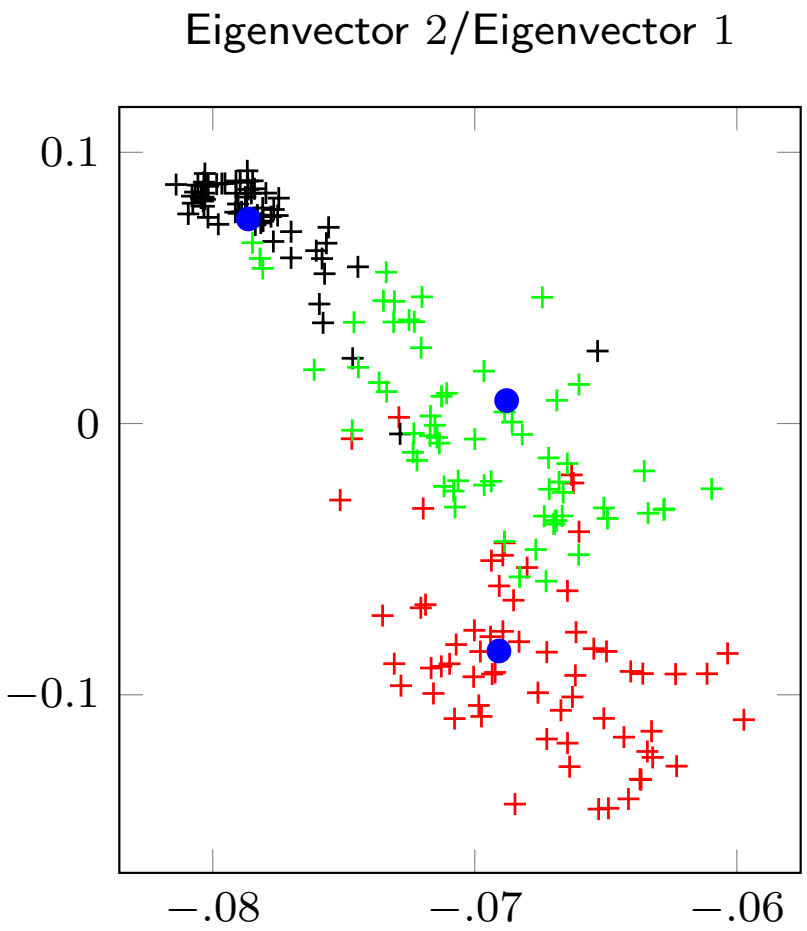
: MNIST data : four leading eigenvectors of $L$ (**red**), versus $\hat{L}$ (**black**) and theoretical findings (**blue**).

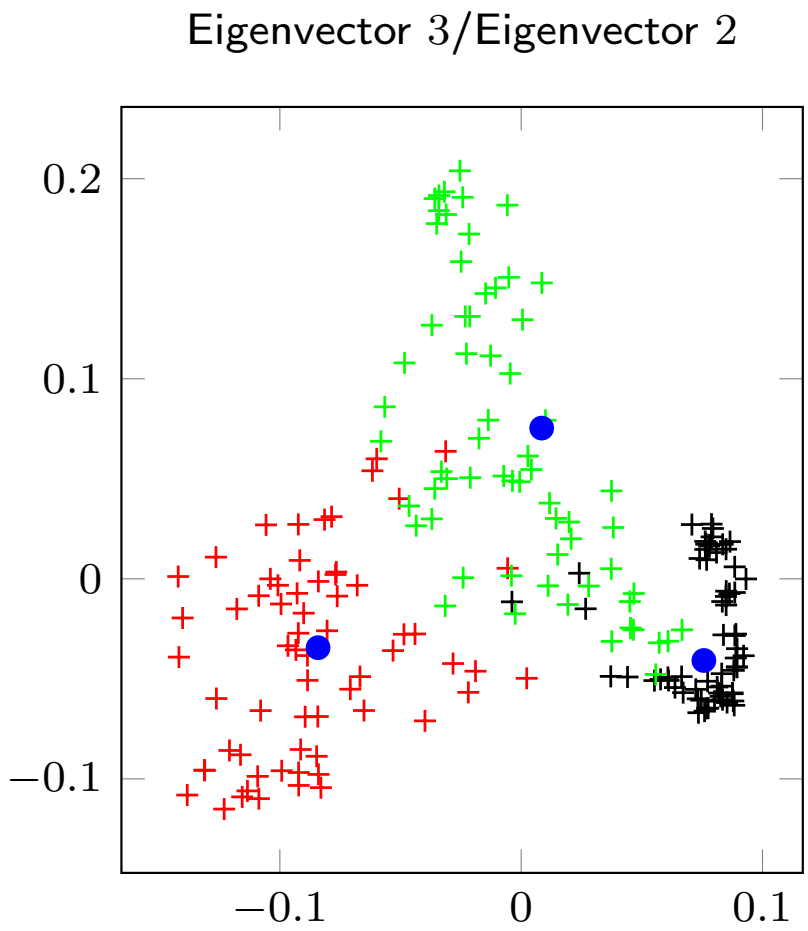
# Class-wise eigenvector means and fluctuations



: MNIST data : four leading eigenvectors of $L$ (**red**), versus $\hat{L}$ (**black**) and theoretical findings (**blue**).

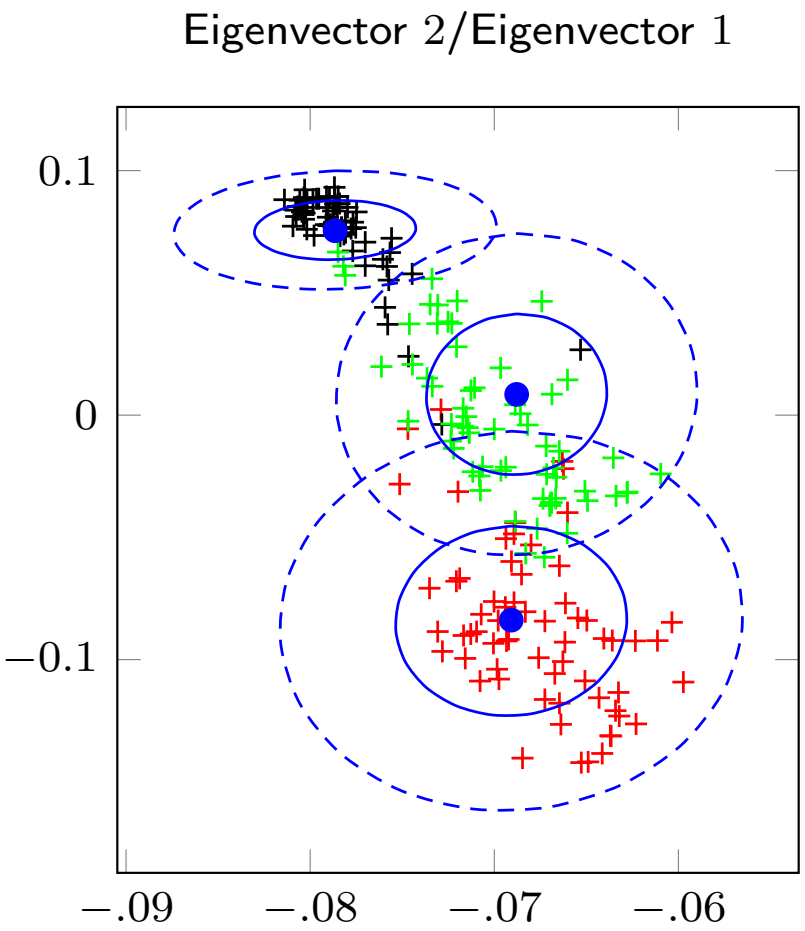
# Class-wise means, fluctuations and cross correlations

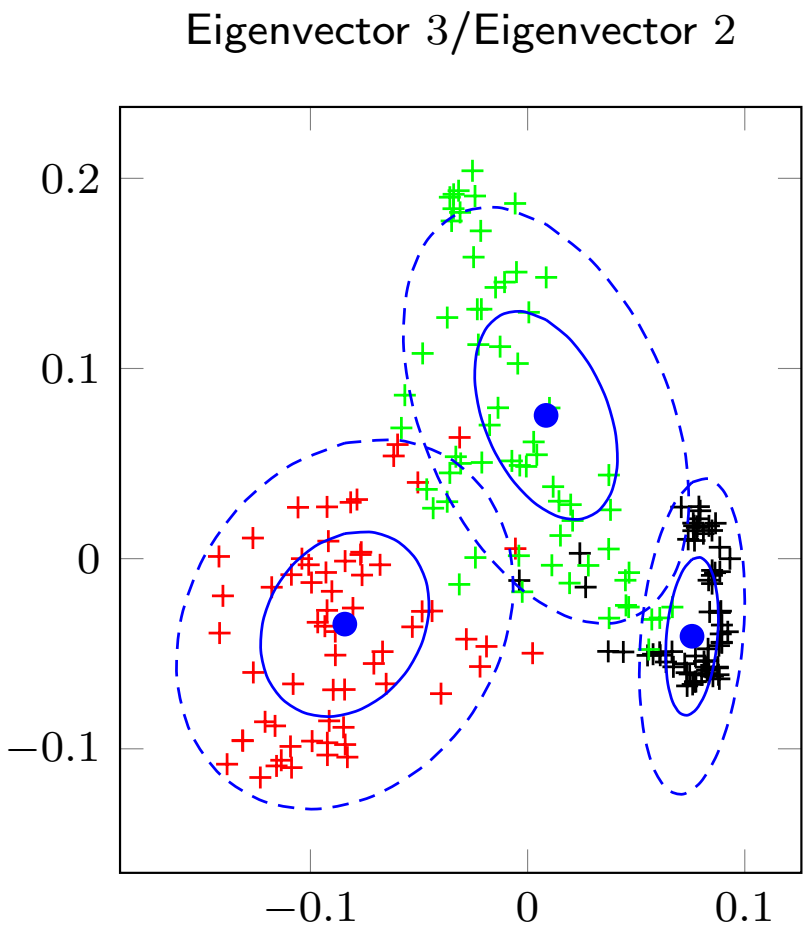



: 2D representation of eigenvectors of $L$, for the MNIST dataset. Theoretical means and 1- and 2-standard deviations in **blue**. Class 1 in **red**, Class 2 in **black**, Class 3 in **green**.

# Class-wise means, fluctuations and cross correlations




: 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**.

# Concluding Remarks

**Summing up** :

- Although Gaussian-based, adequately mimics real world examples
- Noticeable Results :
  - importance of derivatives of $f$ at $\tau$
  - choice of $f(\tau), f'(\tau), f''(\tau)$ determines importance of means, covariances
  - eigenvector may or may not contain information (upon separability condition !)
  - number of isolated eigenvalues not obvious

**Perspectives** :

- (joint) class-wise eigenvector fluctuations
- implications to spectral clustering performance
- algorithm comparison
- ideally, (data-driven) algorithm improvement.