Tomorrow's large dimensional AI: renewed intuitions and new mathematics?

Workshop MACS COMET-SCA on "Automatics and AI"

Romain COUILLET

CentraleSupélec, L2S, University of ParisSaclay, France
GSTATS IDEX DataScience Chair, GIPSA-lab, University Grenoble–Alpes, France.

June 2, 2021

The curse of dimensionality and its consequences

Clustering setting in (not so) large n, p:

The curse of dimensionality and its consequences

Clustering setting in (not so) large n, p:

- GMM setting: $x_1^{(a)}, \ldots, x_{n_a}^{(a)} \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$

The curse of dimensionality and its consequences

Clustering setting in (not so) large n, p:

- GMM setting: $x_1^{(a)}, \ldots, x_{n_a}^{(a)} \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$
- Non-trivial task:

$$\|\mu_a - \mu_b\| = O(1), \quad \operatorname{tr}(C_a - C_b) = O(\sqrt{p}), \quad \operatorname{tr}\left[(C_a - C_b)^2\right] = O(p)$$

The curse of dimensionality and its consequences

Clustering setting in (not so) large n, p:

- GMM setting: $x_1^{(a)}, \ldots, x_{n_a}^{(a)} \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$
- Non-trivial task:

$$\|\mu_a - \mu_b\| = O(1), \quad \operatorname{tr}(C_a - C_b) = O(\sqrt{p}), \quad \operatorname{tr}[(C_a - C_b)^2] = O(p)$$

Classical method: spectral clustering

The curse of dimensionality and its consequences

Clustering setting in (not so) large n, p:

- GMM setting: $x_1^{(a)}, \ldots, x_{n_a}^{(a)} \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$
- Non-trivial task:

$$\|\mu_a - \mu_b\| = O(1), \quad \operatorname{tr}(C_a - C_b) = O(\sqrt{p}), \quad \operatorname{tr}[(C_a - C_b)^2] = O(p)$$

Classical method: spectral clustering

- Extract and cluster the dominant eigenvectors of

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

The curse of dimensionality and its consequences

Clustering setting in (not so) large n, p:

- GMM setting: $x_1^{(a)}, \ldots, x_{n_a}^{(a)} \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$
- Non-trivial task:

$$\|\mu_a - \mu_b\| = O(1), \quad \operatorname{tr}(C_a - C_b) = O(\sqrt{p}), \quad \operatorname{tr}[(C_a - C_b)^2] = O(p)$$

Classical method: spectral clustering

- Extract and cluster the dominant eigenvectors of

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

The curse of dimensionality and its consequences

Clustering setting in (not so) large n, p:

- GMM setting: $x_1^{(a)}, \ldots, x_{n_a}^{(a)} \sim \mathcal{N}(\mu_a, C_a)$, $a = 1, \ldots, k$
- Non-trivial task:

$$\|\mu_a - \mu_b\| = O(1), \quad \operatorname{tr}(C_a - C_b) = O(\sqrt{p}), \quad \operatorname{tr}[(C_a - C_b)^2] = O(p)$$

Classical method: spectral clustering

- Extract and cluster the dominant eigenvectors of

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

- Why? **Finite-dimensional intuition**

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

The curse of dimensionality and its consequences (2)

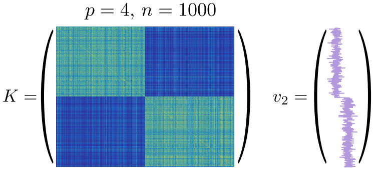
In reality, here is what happens...

Kernel $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$ and second eigenvector v_2 ($x_i \sim \mathcal{N}(\pm\mu, I_p)$, $\mu = (2, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^p$).

The curse of dimensionality and its consequences (2)

In reality, here is what happens...

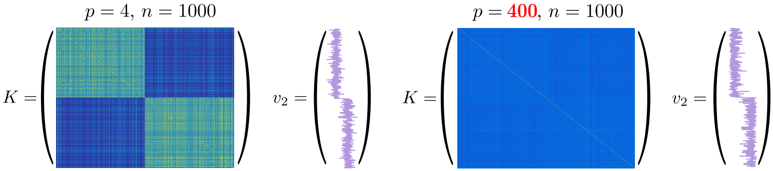
Kernel $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$ and second eigenvector v_2 ($x_i \sim \mathcal{N}(\pm\mu, I_p)$, $\mu = (2, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^p$).



The curse of dimensionality and its consequences (2)

In reality, here is what happens...

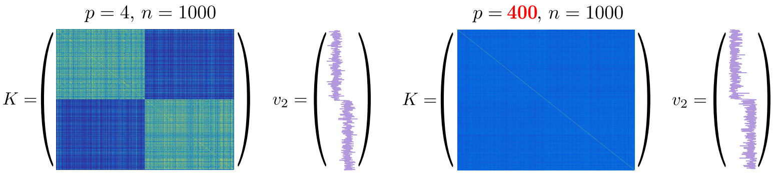
Kernel $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$ and second eigenvector v_2 ($x_i \sim \mathcal{N}(\pm\mu, I_p)$, $\mu = (2, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^p$).



The curse of dimensionality and its consequences (2)

In reality, here is what happens...

Kernel $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$ and second eigenvector v_2
($x_i \sim \mathcal{N}(\pm\mu, I_p)$, $\mu = (2, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^p$).



$p = 4$, $n = 1000$

$p = \mathbf{400}$, $n = 1000$

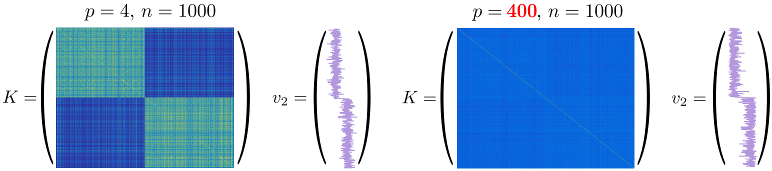
Key observation: Under growth rate assumptions,

$$\boxed{\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p} \|x_i - x_j\|^2 - \boldsymbol{\tau} \right| \right\} \xrightarrow{\text{a.s.}} 0}, \quad \boldsymbol{\tau} = \frac{2}{p} \sum_{i=1}^{k} \operatorname{tr} \frac{n_a}{n} C_a.$$

The curse of dimensionality and its consequences (2)

In reality, here is what happens...

Kernel $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$ and second eigenvector v_2 ($x_i \sim \mathcal{N}(\pm\mu, I_p)$, $\mu = (2, 0, \ldots, 0)^\mathsf{T} \in \mathbb{R}^p$).



$p = 4$, $n = 1000$

$p = \mathbf{400}$, $n = 1000$

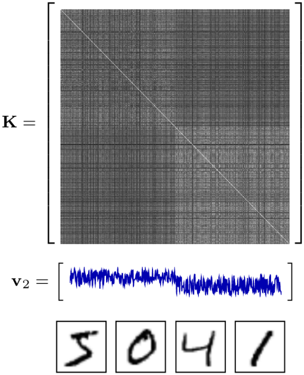
Key observation: Under growth rate assumptions,

$$\boxed{\max_{1 \leq i \neq j \leq n} \left\{ \left| \frac{1}{p}\|x_i - x_j\|^2 - \boldsymbol{\tau} \right| \right\} \xrightarrow{\text{a.s.}} 0}, \quad \boldsymbol{\tau} = \frac{2}{p}\sum_{i=1}^{k} \operatorname{tr} \frac{n_a}{n} C_a.$$

- this suggests $K \simeq f(\boldsymbol{\tau}) 1_n 1_n^\mathsf{T}$!

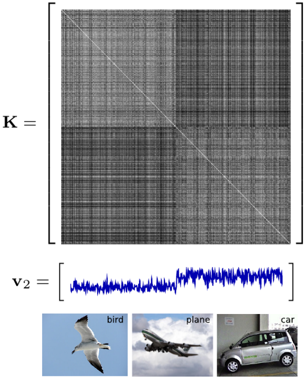
The curse of dimensionality and its consequences (3)

MNIST
raw
$p = 784,\ n = 500$
$\downarrow$



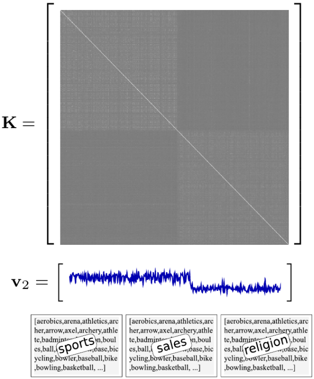
(ici, classes “5” et “0”)

ImageNet
VGG-features
$p = 3084,\ n = 500$
$\downarrow$



(ici, classes “bird” et “plane”)

20NewsGroup
BERT embedding
$p = 300,\ n = 500$
$\downarrow$



(ici, classes “sports” et “sales”)

The curse of dimensionality and its consequences (4)

(Major) consequences:

- Most **machine learning intuitions collapse**

The curse of dimensionality and its consequences (4)

(Major) consequences:

- Most **machine learning intuitions collapse**
- **But luckily**, concentration of distances allows for Taylor expansion, linearization...

The curse of dimensionality and its consequences (4)

(Major) consequences:

- Most **machine learning intuitions collapse**
- **But luckily**, concentration of distances allows for Taylor expansion, linearization...

Theorem (**[C-Benaych'16]** Asymptotic Kernel Behavior)

Under growth rate assumptions, as $p, n \to \infty$,

$$\left\| K - \hat{K} \right\| \xrightarrow{\text{a.s.}} 0, \quad \hat{K} \simeq \underbrace{f(\boldsymbol{\tau}) 1_n 1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)}$$

The curse of dimensionality and its consequences (4)

(Major) consequences:

- Most **machine learning intuitions collapse**
- **But luckily**, concentration of distances allows for Taylor expansion, linearization...

Theorem (**[C-Benaych'16]** Asymptotic Kernel Behavior)

Under growth rate assumptions, as $p, n \to \infty$,

$$\left\| K - \hat{K} \right\| \xrightarrow{\text{a.s.}} 0, \quad \hat{K} \simeq \underbrace{f(\boldsymbol{\tau}) 1_n 1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)} + f'(\boldsymbol{\tau}) \frac{1}{p} Z Z^\mathsf{T} + JAJ^\mathsf{T} + *$$

The curse of dimensionality and its consequences (4)

(Major) consequences:

- Most **machine learning intuitions collapse**
- **But luckily**, concentration of distances allows for Taylor expansion, linearization...

Theorem (**[C-Benaych'16]** Asymptotic Kernel Behavior)

Under growth rate assumptions, as $p, n \to \infty$,

$$\left\| K - \hat{K} \right\| \xrightarrow{\text{a.s.}} 0, \quad \hat{K} \simeq \underbrace{f(\boldsymbol{\tau}) 1_n 1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)} + f'(\boldsymbol{\tau}) \frac{1}{p} Z Z^\mathsf{T} + JAJ^\mathsf{T} + *$$

with $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}$, $j_a = (0, 1_{n_a}, 0)^\mathsf{T}$ *(the clusters!)*

The curse of dimensionality and its consequences (4)

(Major) consequences:

- Most **machine learning intuitions collapse**
- **But luckily**, concentration of distances allows for Taylor expansion, linearization...

Theorem (**[C-Benaych'16]** Asymptotic Kernel Behavior)

Under growth rate assumptions, as $p, n \to \infty$,

$$\left\| K - \hat{K} \right\| \xrightarrow{\text{a.s.}} 0, \quad \hat{K} \simeq \underbrace{f(\boldsymbol{\tau}) 1_n 1_n^{\mathsf{T}}}_{O_{\|\cdot\|}(n)} + f'(\boldsymbol{\tau}) \frac{1}{p} Z Z^{\mathsf{T}} + J A J^{\mathsf{T}} + *$$

with $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}$, $j_a = (0, 1_{n_a}, 0)^{\mathsf{T}}$ *(the clusters!) and* $A \in \mathbb{R}^{k \times k}$ *function of:*

- $f(\boldsymbol{\tau})$, $f'(\boldsymbol{\tau})$, $f''(\boldsymbol{\tau})$
- $\|\mu_a - \mu_b\|$, *tr* $(C_a - C_b)$, *tr* $((C_a - C_b)^2)$, *for* $a, b \in \{1, \ldots, k\}$.

The curse of dimensionality and its consequences (4)

(Major) consequences:

- Most **machine learning intuitions collapse**
- **But luckily**, concentration of distances allows for Taylor expansion, linearization...

Theorem ([C-Benaych'16] Asymptotic Kernel Behavior)

Under growth rate assumptions, as $p, n \to \infty$,

$$\left\| K - \hat{K} \right\| \xrightarrow{\text{a.s.}} 0, \quad \hat{K} \simeq \underbrace{f(\boldsymbol{\tau}) 1_n 1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)} + f'(\boldsymbol{\tau}) \frac{1}{p} Z Z^\mathsf{T} + J A J^\mathsf{T} + *$$

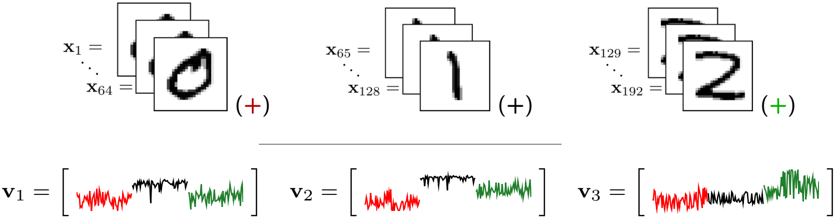
with $J = [j_1, \ldots, j_k] \in \mathbb{R}^{n \times k}$, $j_a = (0, 1_{n_a}, 0)^\mathsf{T}$ (the clusters!) and $A \in \mathbb{R}^{k \times k}$ function of:

- $f(\boldsymbol{\tau})$, $f'(\boldsymbol{\tau})$, $f''(\boldsymbol{\tau})$
- $\|\mu_a - \mu_b\|$, *tr* $(C_a - C_b)$, *tr* $((C_a - C_b)^2)$, *for* $a, b \in \{1, \ldots, k\}$.

⇒ **This is a spiked model! We can study it fully!**

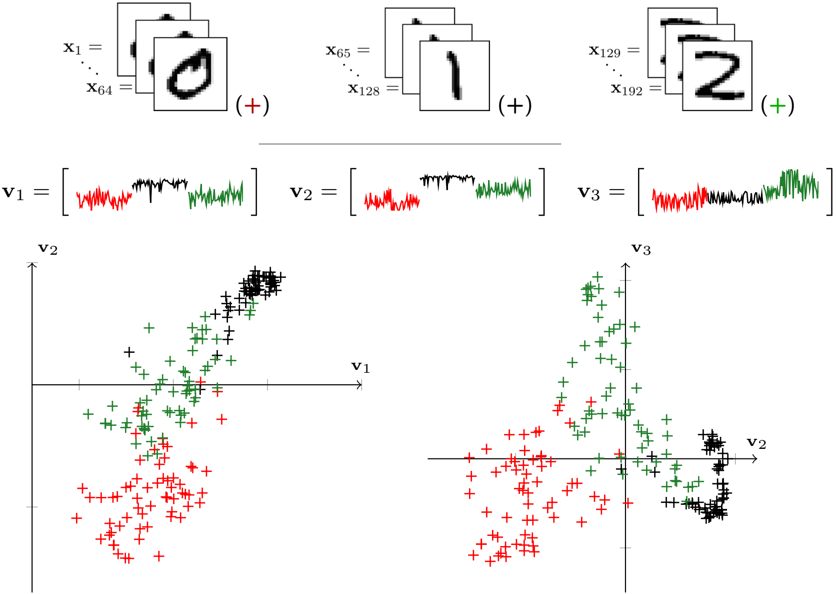
Performance prediction: spectral clustering

- **Asymptotic analysis of eigenvectors of** K: (MNIST, $p = 28 \times 28 (= 784)$)



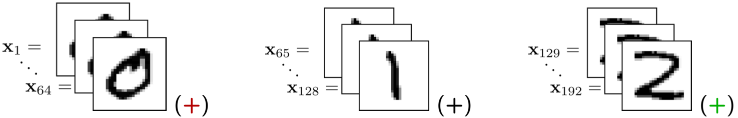
Performance prediction: spectral clustering

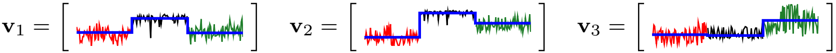
- **Asymptotic analysis of eigenvectors of** K: (MNIST, $p = 28 \times 28 (= 784)$)

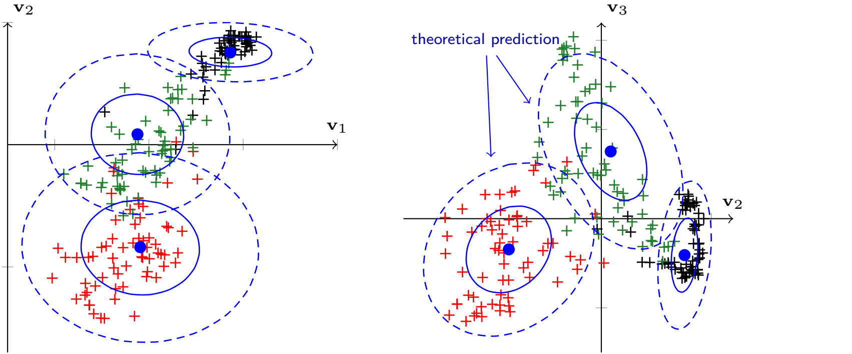


Performance prediction: spectral clustering

- **Asymptotic analysis of eigenvectors of** K: (MNIST, $p = 28 \times 28 (= 784)$)







theoretical prediction

Takeaway Message 2

“RMT Reassesses and Improves Data Processing”

Today's menu:

1. New Counter-Intuitive Kernels
2. Resurrecting Semi-Supervised Learning
3. Making complex ML frameworks simple: Multitask Learning
4. Towards cheap "environment-friendly" learning
5. Using Random Matrices to Study... Random Tensors!

1. New Counter-Intuitive Kernels

Improving Kernel Spectral Clustering

- **Going further than ([Kammoun,Couillet'17])**,

$$K \simeq \underbrace{f(\tau)1_n 1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)} + f'(\tau)\frac{1}{p}ZZ^\mathsf{T} + JAJ^\mathsf{T}, \text{ avec } A = F\begin{pmatrix} f(\tau), f'(\tau), f''(\tau) \\ \|\mu_a - \mu_b\|, \operatorname{tr}(C_a - C_b), \ldots \end{pmatrix}.$$

Improving Kernel Spectral Clustering

- **Going further than ([Kammoun,Couillet'17])**, if $\underline{f'(\tau) = 0}$,

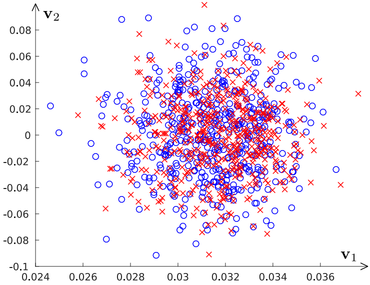
$$K \simeq \underbrace{f(\tau)1_n1_n^{\mathsf{T}}}_{O_{\|\cdot\|}(n)} + \cancel{f'(\tau)\frac{1}{p}ZZ^{\mathsf{T}}} + JAJ^{\mathsf{T}}, \text{ avec } A = F\begin{pmatrix} f(\tau), f'(\tau), f''(\tau) \\ \cancel{\|\mu_a - \mu_b\|}, \mathrm{tr}\,(C_a - C_b), \ldots \end{pmatrix}.$$

Improving Kernel Spectral Clustering

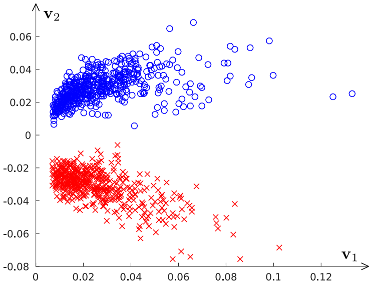
- **Going further than ([Kammoun,Couillet'17])**, if $\underline{f'(\tau)=0}$,

$$K \simeq \underbrace{f(\tau)1_n1_n^\mathsf{T}}_{O_{\|\cdot\|}(n)} + \cancel{f'(\tau)\frac{1}{p}ZZ^\mathsf{T}} + JAJ^\mathsf{T}, \text{ avec } A = F\left(\begin{array}{l} f(\tau), f'(\tau), f''(\tau) \\ \cancel{\|\mu_a-\mu_b\|}, \mathrm{tr}\,(C_a-C_b), \ldots \end{array}\right).$$

- **Gaussian case**: $\mathcal{N}(0,\mathbf{C}_1)$ vs. $\mathcal{N}(0,\mathbf{C}_2)$

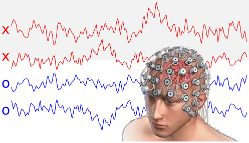


Kernel $K_{ij} = \exp(-\frac{1}{2p}\|x_i-x_j\|^2)$

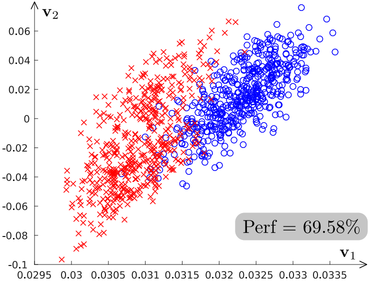


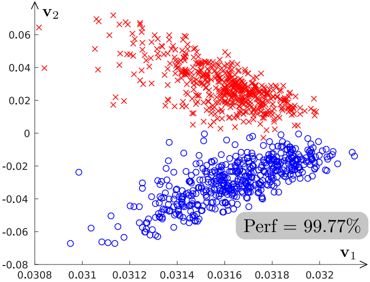
Kernel $K_{ij} = (\frac{1}{p}\|x_i-x_j\|^2 - \tau)^2$

Improving Kernel Spectral Clustering



- **EEG data**: sane vs. epileptic patients

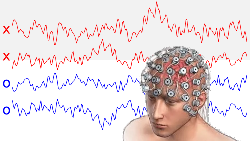




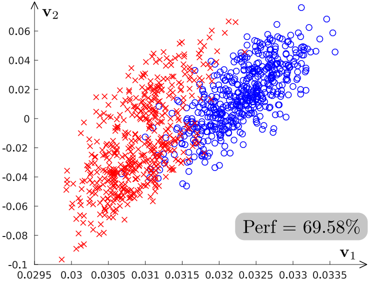
Kernel $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$

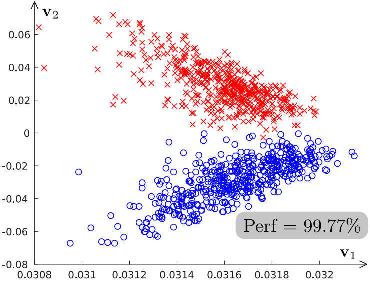
Kernel $K_{ij} = (\frac{1}{p}\|x_i - x_j\|^2 - \tau)^2$

Improving Kernel Spectral Clustering



- **EEG data**: sane vs. epileptic patients





Kernel $K_{ij} = \exp(-\frac{1}{2p}\|x_i - x_j\|^2)$

Kernel $K_{ij} = (\frac{1}{p}\|x_i - x_j\|^2 - \tau)^2$

→ ***Remark***: *highly counter-intuitive kernel!*