# Kernel-based Methods for Unsupervised Learning

LEAR project-team, INRIA

Zaid Harchaoui

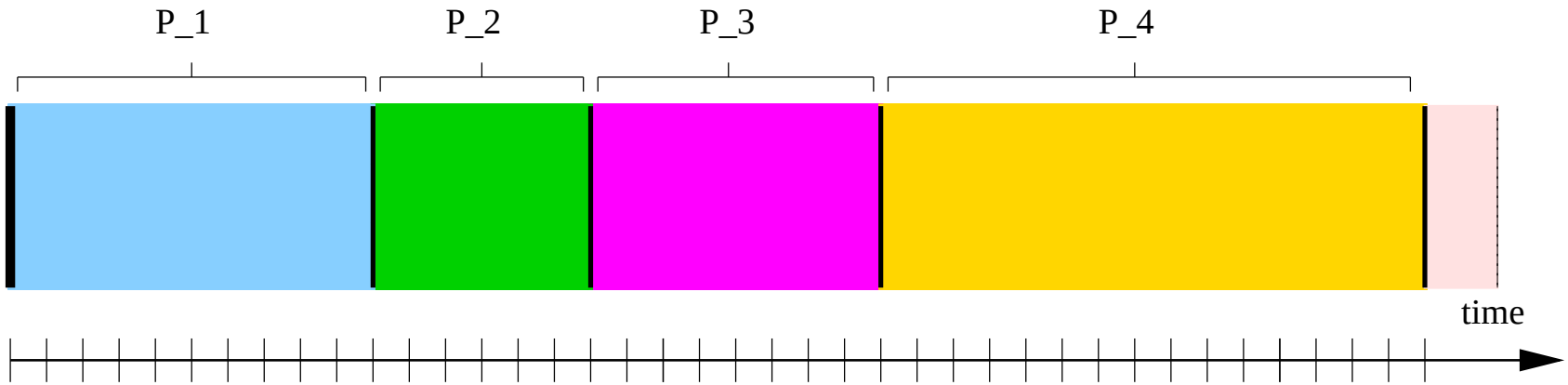
Lyon, Janvier 2011

# Temporal segmentation (clustering with temporal consistency)

## Change-in-mean model

Time series of independent r.v. $\{Y_t\}_{t=1,\dots,n}$ such that

$$Y_t \overset{\mathcal{D}}{\sim} \mathcal{N}(\mu_k^\star, \sigma^2)\ , \quad t_{k-1}^\star + 1 \leq t \leq t_k^\star, \quad k = 1, \dots, K^\star + 1\ , \qquad (1)$$

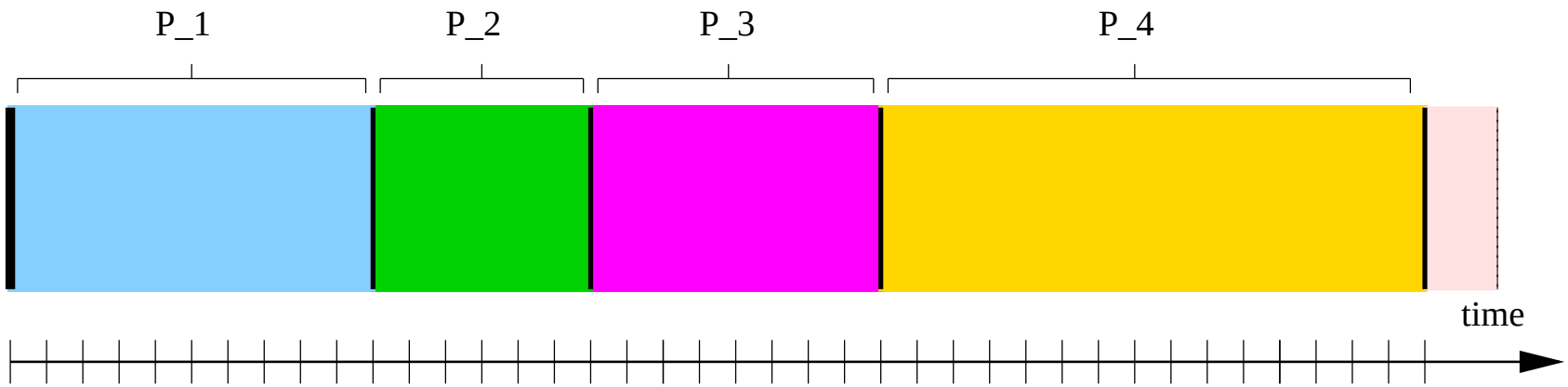


# Temporal segmentation

## Change-in-mean-element model

Time series of independent r.v. $\{Y_t\}_{t=1,\dots,n}$ such that

$$\mathbb{E}[k(Y_t,\cdot)] = \mu_k^\star\ , \quad t_{k-1}^\star + 1 \leq t \leq t_k^\star, \quad k = 1,\dots,K^\star+1\ .$$



# Temporal segmentation with kernels

Classical least-squares formulation

$$\underset{t_1,\ldots,t_{K^\star}}{\text{Minimize}} \sum_{k=1}^{K^\star+1} \sum_{t=t_{k-1}+1}^{t_k} (Y_t - \overline{Y}(t_{k-1}, t_k))^2$$

Kernel-based version in $\mathcal{H}$

$$\underset{t_1,\ldots,t_{K^\star}}{\text{Minimize}} \sum_{k=1}^{K^\star+1} \sum_{t=t_{k-1}+1}^{t_k} \left\| k(Y_t, \cdot) - \hat{\mu}_{[t_{k-1}:t_k]} \right\|_{\mathcal{H}}^2$$

# Massaging the objective function

## Intra-segment scatter

$$\underset{t_1,\ldots,t_{K^\star}}{\text{Minimize}} \sum_{k=1}^{K-1} \hat{V}(Y_{t_k+1},\ldots,Y_{t_{k+1}})$$

with $\hat{V}(Y_{t+1},\ldots,Y_{t+s}) = \left\| k(Y_t,\cdot) - \hat{\mu}_{[t+1:t+s]} \right\|_{\mathcal{H}}^2$

# Forward-backwrad recursions

## Forward recursions

$$
\begin{aligned}
I_k(t) &= \operatorname*{Min}_{t_1,\ldots,t_{k-1};t_k=t} \sum_{k=1}^{K-1} \hat{V}(Y_{t_k+1},\ldots,Y_{t_{k+1}}) \\
&= \operatorname*{Min}_{t_{k-1};t_k=t} \operatorname*{Min}_{t_1,\ldots,t_{k-2}} \sum_{k=1}^{K-1} \hat{V}(Y_{t_k+1},\ldots,Y_{t_{k+1}}) \\
&= \operatorname*{Min}_{t_{k-1}} (I_{k-1}(t_{k-1}) + \hat{V}(Y_{t_{k-1}},\ldots,Y_t)) \ .
\end{aligned}
$$

## Dynamic programming

Dynamic programming algorithm working on submatrices of the Gram matrix, leading to a time-complexity of $O(Kn^2)$.

# Mental task segmentation

## Dataset

- Data : $3$ normal subjects during $4$ non-feedback sessions
- $3$ tasks : imagination of repetitive self-paced left hand movements or right hand movements, and generation of words beginning with the same random letter
- Features : based on Power Spectral Density

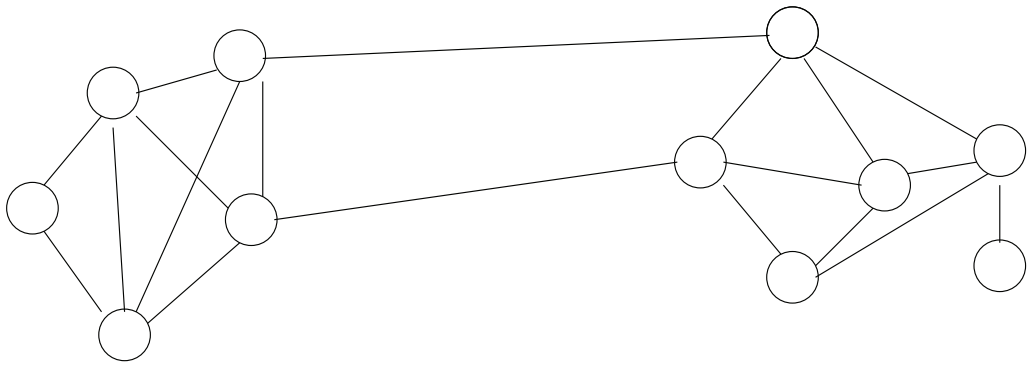
## Experimental results

| | Subject 1 | Subject 2 | Subject 3 |
|---|---|---|---|
| KCpA | 79% | 74% | 61% |
| SVM | 76% | 69% | 60% |

# Spectral clustering (von Luxburg, 2007)

## Overview

- Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ a dataset of points in $\mathbf{R}^d$, along with pairwise similarities $s(\mathbf{x}_i, \mathbf{x}_j), 1 \leq i, j \leq n$.
- Build similarity graph, with data points as *vertices* and similarities as *edge lengths*
- Spectral clustering finds the best cut through the graph



# Laplacian matrix and spectral clustering

## Laplacian matrix

Spectral clustering relies on the spectrum of the Laplacian matrix $\mathbf{L}$

$$\mathbf{L} = \underbrace{\mathbf{D}}_{\text{degree matrix}} - \underbrace{\mathbf{S}}_{\text{similarity matrix}} ,$$

where

$$\mathbf{D} = \text{Diag}(\deg(\mathbf{x}_1), \ldots, \deg(\mathbf{x}_n))$$

$$\deg(\mathbf{x}_i) = \sum_{j=1}^{n} s(\mathbf{x}_i, \mathbf{x}_j) \ .$$

# Laplacian matrix and the Laplace-Beltrami operator

## Laplacian matrix

The Laplacian matrix measures the discrete variation of $f$ along the graph

$$\forall f \in \mathbb{R}^d \, , f^T \mathbf{L} f = \frac{1}{2} \sum_{j=1}^{n} s(\mathbf{x}_i, \mathbf{x}_j)(f_i - f_j)^2 \, ,$$

$$f^T \mathbf{L} f \approx \frac{1}{2} \sum_{j=1}^{n} \frac{(f_i - f_j)^2}{d(\mathbf{x}_i, \mathbf{x}_j)^2} \, , \quad \text{if } s(\mathbf{x}_i, \mathbf{x}_j) \approx \frac{1}{d(\mathbf{x}_i, \mathbf{x}_j)^2} \, .$$

## Laplacian operator

The Laplacian matrix is the discrete counterpart of the Laplace[1] operator

$$\forall f \in \mathbb{R}^d \, , \langle f, \Delta f \rangle = \int |\nabla f|^2 dx \, .$$

1. Laplace-Beltrami generalizes the Laplace operator to manifold data.

# Rescue theorems

## Properties of Laplacian operators

Laplacian matrices are (von Luxburg et al., 2005, Gine and Koltchinskii, 2008)

- symmetric
- positive definite
- smallest eigenvalue is $0$, and associated eigenvector $\mathbf{1}$

## Interpretation

- Multiplicity of eigenvalue $0$ is the number of connected components of the graph $A_1, \ldots, A_k$
- Eigenspace spanned by the characteristic functions $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}$ on those components (so all eigenvectors are piecewise constants)

# Normalization

## Normalized graph Laplacians

Graph Laplacian matrices can be normalized in two ways [2]

$$\mathbf{L}_{rw} = D^{-1}L \qquad \text{random walk normalization ,}$$
$$\mathbf{L}_{sym} = D^{-1/2}LD^{-1/2} \qquad \text{symmetrized normalization .}$$

## Interpretation

- $\mathbf{L}_{rw}$ and $\mathbf{L}_{sym}$ share similar spectral properties with $\Lambda$
- Normalized graph Laplacians are better understood theoretically and are consistent under general assumptions in large-sample settings
- Un-normalized ones are still used ( !) despite their lack of consistency in some cases in large-sample settings.

2. Caution : eigenspace of $\mathbf{L}_{rw}$ spanned by the $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}$ ; eigenspace of $\mathbf{L}_{sym}$ spanned by the $D^{1/2}\mathbf{1}_{A_1}, \ldots, D^{1/2}\mathbf{1}_{A_k}$.

# Spectral clustering

## Spectral clustering algorithm

- Build similarity graph
- Performs an SVD on $\mathbf{L}_{rw}$ or $\mathbf{L}_{sym}$ to get the first $k$ eigenvector/eigenvalue pairs $(v_j, \lambda_j)_{j=1,\ldots,c}$.
- Build the matrix $V = [v_1, \ldots, v_k]$ stacking the $k$ eigenvectors as columns
- Launch your *favourite clustering algorithm* on the $n$ rows of $V$

# Reminder : $k$-means (a.k.a Lloyd's quantization algorithm)

## The $k$-means objective function

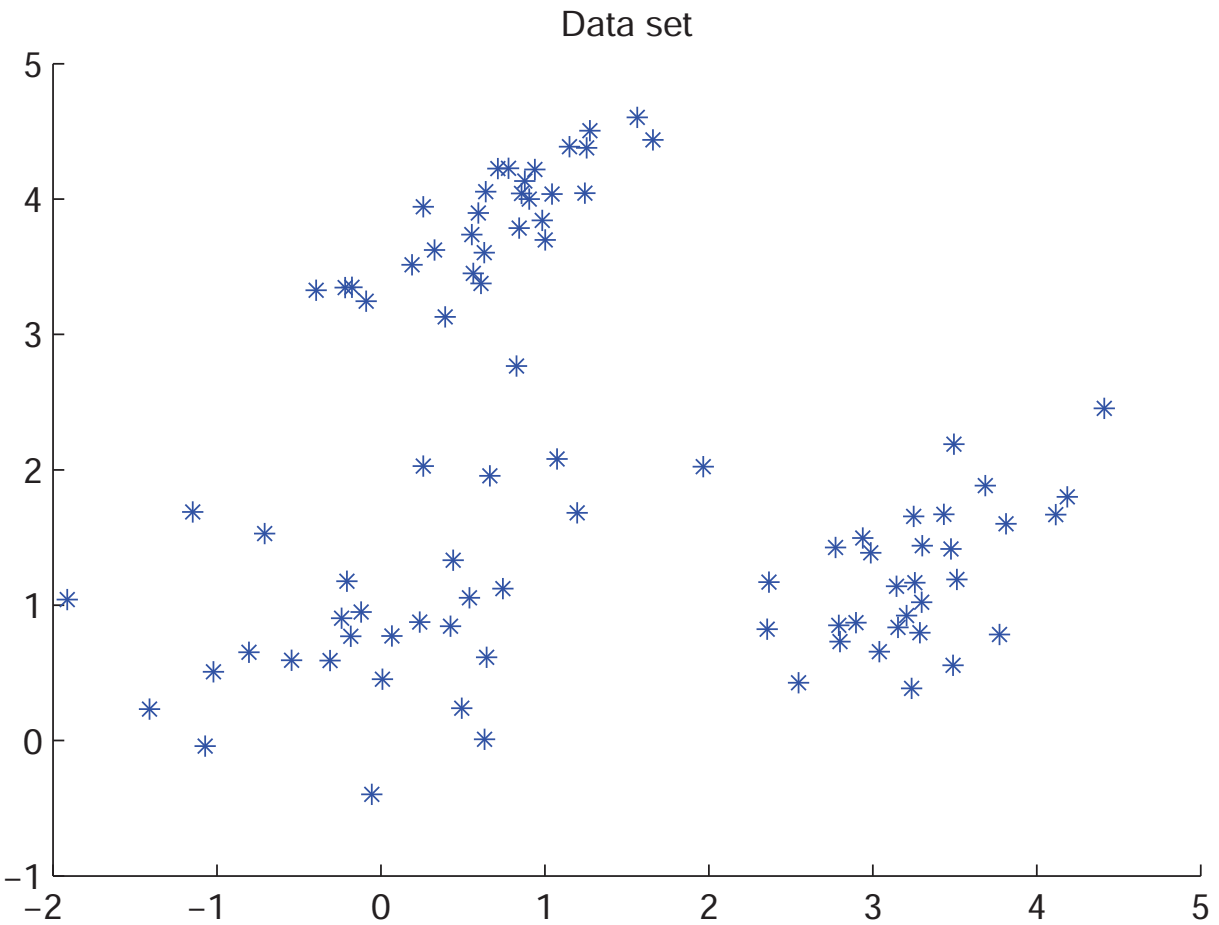
- Goal : given $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, find a partition of the $n$ observations into $k$ sets $S_1, \ldots, S_k$ so as to minimize $\sum_{\ell=1}^{k} \sum_{\mathbf{x}_j \in S_\ell} \|\mathbf{x}_j - \mu_\ell\|^2$, with $\mu_\ell$ is the mean over $S_\ell$

## The $k$-means algorithm

- Initialization : initial set of means
- Assignment : assign each observation to the cluster with the closest mean
- Update : compute each new mean to be the centroid (isobarycenter) of the observations in the cluster

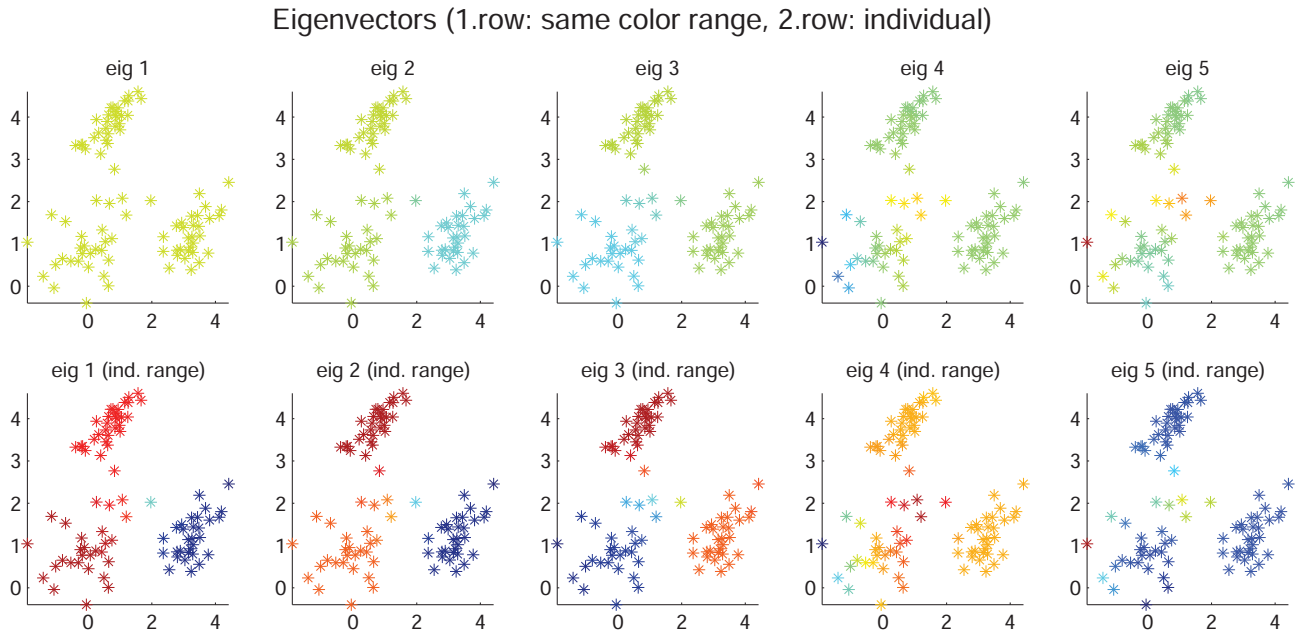
# Example

## 2D example with 3 clusters



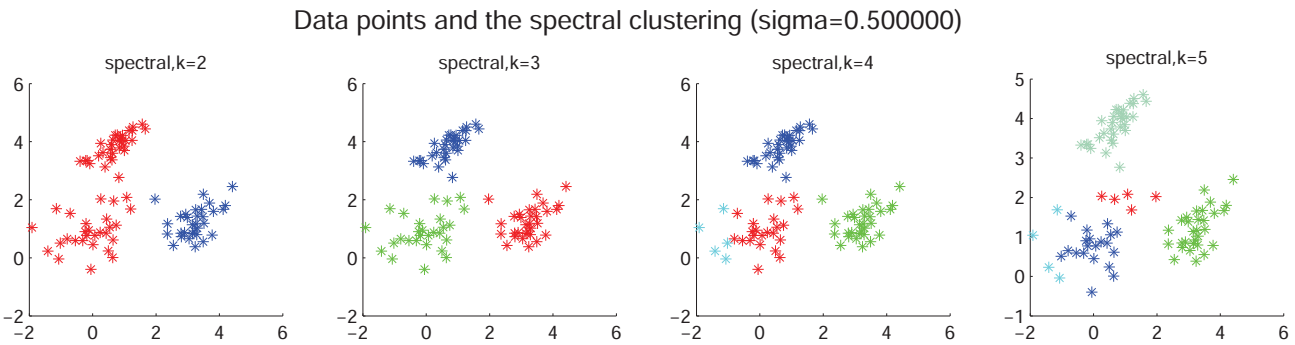
# Example

## Projections onto eigenvectors



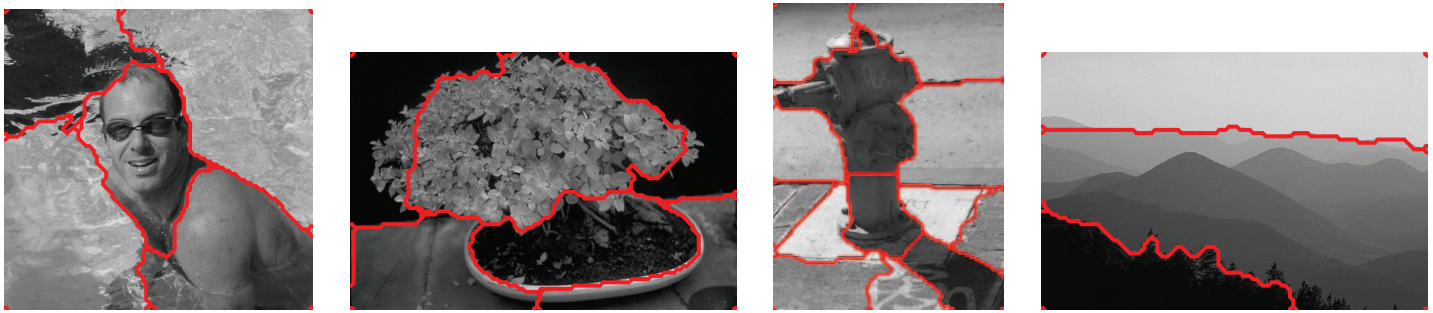
# Example

## Clustering obtained with $k$-means as the favourite clustering algorithm



# Spectral clustering for image segmentation

## Image segmentation algorithm



# GrabCut and foreground extraction

## Interactive foreground extraction algorithm

# Testing for homogeneity

## Homogeneity of two samples

- Two samples $X_1^{(1)}, \ldots, X_{n_1}^{(1)} \sim \mathbb{P}^{(1)}$ and $X_1^{(2)}, \ldots, X_{n_2}^{(2)} \sim \mathbb{P}^{(2)}$ independent and mutually independent
- Problem : decide between

$$\mathbf{H}_0 : \quad \mathbb{P}^{(1)} = \mathbb{P}^{(2)}$$
$$\mathbf{H}_A : \quad \mathbb{P}^{(1)} \neq \mathbb{P}^{(2)}$$

# Reminder on statistical hypothesis testing

## Decision rule

$$\text{If} \quad \underbrace{T_n}_{\text{test statistic}} > \underbrace{c_{1-\alpha}}_{\text{significance level}} \quad \text{then decide } \mathbf{H}_A$$

## Type I and Type II errors

**Significance level $\alpha$**
(Type I error probability) :

$$\mathbb{P}_{\mathbf{H}_0}(\text{decide } \mathbf{H}_A) \leq \alpha$$

**Power against $h$ $\pi(h)$**
(1 -Type II error prob.)

$$\forall h \in \mathbf{H}_A \ , \quad \pi(h) = \mathbb{P}_h(\text{decide } \mathbf{H}_A)$$

## in our setting

Asymptotic level $\alpha$
$\hookrightarrow$ large-sample distribution under $\mathbf{H}_0$, as $n \to \infty$

Asymptotic power 1
$\hookrightarrow$ large-sample distribution under $\mathbf{H}_A$, as $n \to \infty$

# Test statistic

Empirical mean elements $\hat{\mu}_1$ and $\hat{\mu}_2$, and Empirical covariance operators $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ resp. $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\}$ et $\{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$

$$\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}\} \hookrightarrow (\hat{\mu}_1, \hat{\Sigma}_1) \quad \text{and} \quad \{X_1^{(2)}, \ldots, X_{n_2}^{(2)}\} \hookrightarrow (\hat{\mu}_2, \hat{\Sigma}_2) \; .$$

## Regularized Kernel Fisher Discriminant Ratio

$$\begin{aligned} &\mathrm{KFDR}_{n_1,n_2;\gamma}(X_1^{(1)}, \ldots, X_{n_1}^{(1)}; X_1^{(2)}, \ldots, X_{n_2}^{(2)}) \\ &\stackrel{\text{def}}{=} \frac{n_1 n_2}{n_1 + n_2} \left\| \left( \underbrace{\frac{n_1}{n}\hat{\Sigma}_1 + \frac{n_2}{n}\hat{\Sigma}_2}_{\hat{\Sigma}_W} + \gamma \mathrm{I} \right)^{-1/2} (\hat{\mu}_2 - \hat{\mu}_1) \right\|_{\mathcal{H}}^2 . \end{aligned}$$

Hotelling's $T^2$ : homogeneity of two normal distributions in the mean with equal but unknown covariances

$$\frac{n_1 n_2}{n_1 + n_2} \left\| \left( \frac{n_1}{n}\hat{\Sigma}_1 + \frac{n_2}{n}\hat{\Sigma}_2 \right)^{-1/2} (\hat{\mu}_2 - \hat{\mu}_1) \right\|_{\mathfrak{R}^d}^2$$

# Limiting distribution under $H_0 : \gamma_n \equiv \gamma$

## Proposition

Assume the kernel is bounded and that for $a = 1, 2$ the eigenvalues $\{\lambda_p(\Sigma_a)\}_{p\geq 1}$ satisfy $\sum_{p=1}^{\infty} \lambda_p^{1/2}(\Sigma_a) < \infty$. Assume in addition that $\mathbb{P}_1$ and $\mathbb{P}_2$ are equal *i.e.* $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$, and $\gamma_n \equiv \gamma > 0$. Then,

$$\frac{\mathrm{KFDR}_{n_1,n_2;\gamma} - d_{1,n_1,n_2;\gamma}(\hat{\Sigma}^W_{n_1,n_2})}{\sqrt{2}\, d_{2,n_1,n_2;\gamma}(\hat{\Sigma}^W_{n_1,n_2})} \xrightarrow{\mathcal{D}} \frac{1}{\sqrt{2}\, d_{2,n_1,n_2;\gamma}(\Sigma_W)} \sum_{p=1}^{\infty} \frac{\lambda_p(\Sigma_W)}{\lambda_p(\Sigma_W) + \gamma} (\underbrace{Z_p^2}_{\chi_1^2} - 1) \,,$$

## Remarks

$$d_{1,n_1,n_2;\gamma}(\hat{\Sigma}_W) \stackrel{\mathrm{def}}{=} \mathrm{Tr}((\hat{\Sigma}_W + \gamma I)^{-1} \hat{\Sigma}_W) \qquad \text{recentering}$$

$$d_{2,n_1,n_2;\gamma}(\hat{\Sigma}_W) \stackrel{\mathrm{def}}{=} [\mathrm{Tr}((\hat{\Sigma}_W + \gamma I)^{-2} \hat{\Sigma}_W^2)]^{1/2} \qquad \text{renormalisation}$$

# Consistency in power

## Proposition

Assume that the kernel is bounded and that for $a = 1, 2$ the eigenvalues $\{\lambda_p(\Sigma_a)\}_{p\geq 1}$ satisfy $\sum_{p=1}^{\infty} \lambda_p^{1/2}(\Sigma_a) < \infty$, and that the RKHS $\mathcal{H}$ is dense in $L^2(\mathbb{P})$ for all $\mathbb{P}$. Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability distributions such that $\mathbb{P}_2 \neq \mathbb{P}_1$. Then,

$$\mathbb{P}_{\mathbf{H}_A}\left(\frac{\mathrm{KFDR}_{n_1,n_2;\gamma_n} - d_{1,n_1,n_2;\gamma}(\hat{\Sigma}^W_{n_1,n_2})}{\sqrt{2}\, d_{2,n_1,n_2;\gamma}(\hat{\Sigma}^W_{n_1,n_2})} > c_{1-\alpha}\right) \to 1\,. \tag{2}$$

## Remarks

University density satisfied for translation-invariant kernels $k(x, y) = k(x - y)$ such as the gaussian kernel (Steinwart, 2006 ; Sriperumbudur et al., 2008).

# In practice

## Kernel trick

$$\left\|(\hat{\Sigma}_W + \gamma_n \mathrm{I})^{-1/2}(\hat{\mu}_2 - \hat{\mu}_1)\right\|_{\mathcal{H}}^2$$
$$= \gamma^{-1}\left\{\mathbf{m}_n^T \mathbf{K}_n \mathbf{m}_n - n^{-1}\mathbf{m}_n^T \mathbf{K}_n \mathbf{N}_n (\gamma \mathrm{I} + n^{-1}\mathbf{N}_n \mathbf{K}_n \mathbf{N}_n)^{-1}\mathbf{N}_n \mathbf{K}_n \mathbf{m}_n\right\} .$$

$\mathbf{K}_n = [k(x_i, x_j)]_{i,j=1,\ldots,n}$ is the Gram matrix, $\mathbf{N}_n$ is intra-class recentering matrix (each block correspond to one sample), and $\mathbf{m}_n = (\mathbf{m}_{n,i})_{1 \leq i \leq n}$ "difference-in-mean vector" with $\mathbf{m}_{n,i} = -n_1^{-1}$ pour $i = 1, \ldots, n_1$ et $\mathbf{m}_{n,i} = n_2^{-1}$ for $i = n_1 + 1, \ldots, n_1 + n_2$

## Complexity

$O((n_1 + n_2)^2)$ is space (storing Gram matrix) and $O((n_1 + n_2)^3)$ in time (linear system solving)

# Application : speaker verification

- 8 speakers from NIST 2004 evaluation
- features : MFCC

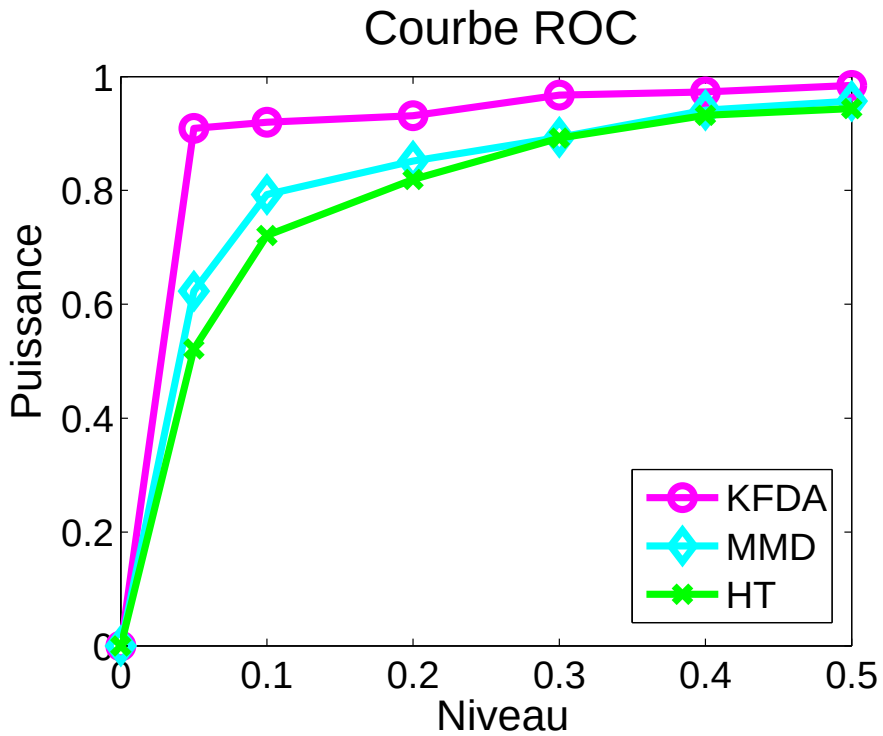



Figure: Comparison of ROC curves

# Application : audio segmentation audio

## TV-shows from 80s "Grand Echiquier"

- Semantic segmentation (global segmentation) : applause/film/music/*interview*
- Speaker segmentation (local segmentation) : Coluche/J. Chancel/F.-R. Duchable/etc.

| | Nb. de sections | Duree moyenne (sec.) |
|---|---|---|
| applaud. | 84 | 14 |
| film | 29 | 155 |
| musique | 38 | 194 |
| parole | 188 | 70 |
| tours de loc. | 962 | 6 |

Table: Description des donnees

# Experimental results in audio segmentation

- sliding window along the signal assuming change occuring in the middle of the window
- super-features built from MFCC
- comparison with unsupervised MMD (Gretton et al., 2004), KCD (Desobry et al., 2005), and supervised HMM (Rabiner et al., 2007)

| | Seg. semantique | | Seg. locuteurs | |
|---|---|---|---|---|
| | Precision | Rappel | Precision | Rappel |
| KFDR | 0.72 | 0.63 | 0.89 | 0.90 |
| MMD | 0.71 | 0.58 | 0.76 | 0.73 |
| KCD | 0.65 | 0.63 | 0.78 | 0.74 |
| HMM | 0.73 | 0.65 | 0.93 | 0.96 |

Table: Precision et Rappel