A study of the manifold hypothesis for functional data by using spectral clustering

Julien Ah-Pine\textsuperscript{1,2} \hspace{1cm} Anne-Françoise Yao\textsuperscript{2}

\textsuperscript{1}Univ. of Lyon - ERIC Lab

\textsuperscript{2}Univ. of Clermont Auvergne - LMBP

CMStatistics 2019
London, 14th of June 2019
Outline

1 Background and Motivations

2 Spectral clustering

3 Experiments and discussion
Outline

1. Background and Motivations
2. Spectral clustering
3. Experiments and discussion
Functional Data (FD)

- In many applications, observations are realization of functional data (FD) (curves, time series, signals, images, . . .).
- Functional Data Analysis (FDA) extends multivariate data analysis techniques to FD or develops specific techniques for FD, see for e.g. [?, ?].
- Objects under study are $n$ real valued functions $\{x_i\}_{i=1,...,n}$ in $L^2([0, T])$, where $T > 0$.
- However $\forall x_i$, we only have $p$ measurements $\{y_{ij}\}_{j=1,...,p}$ at discrete time points $\{t_j\}_{j=1,...,p}$ in $[0, T]$ and these observations are assumed to be corrupted by noise $\epsilon_{ij}$:

$$y_{ij} = x_i(t_j) + \epsilon_{ij}, \quad \forall i, \forall j$$

where $\epsilon_{ij}$ are assumed to be independent across $i$ and $j$. 
Functional Data Clustering (FDC)

- Given \( \{y_{ij}\}_{i,j} \) find a partition of \( \{x_i\}_i \) where FD in a class are more similar to each other than to FD in other classes (see for e.g. [?]).
- One possible workflow for FDC is the following one :
  - Represent the FD in a **low-dimensional space** using either :
    - Pre-defined finite set of basis functions such as bsplines.
    - Data-driven finite set of basis functions such as truncated Karhunen-Loeve expansion (a.k.a. functional PCA).
  - Apply **multivariate clustering techniques** either :
    - Assuming all FD belong to the whole low-dimensional representation space (e.g. \( k \)-means or hierarchical clustering).
    - Assuming that each cluster only belong to a subspace of the representation space (e.g. subspace clustering or model-based functional clustering techniques).
Background and Motivations

Example: Berkeley Growth data

\[ \{y_{ij}\}_{j=1,...,p} = \text{heights measured at different times } t_j. \]

Discrete observations of 2 FD time value

J. Ah-Pine, A.-F. Yao
A study of manifold hypothesis for FD
CMStatistics 2019
Example: Berkeley Growth data

- $\{y_{ij}\}_{j=1}^{p} = \text{heights measured at different times } t_j$.
- $x_i = \text{height function of individual } i$.

**Discrete observations of 2 FD**

**Continuous representation of 2 FD**
Example: Berkeley Growth data

- $\{y_{ij}\}_{j=1,...,p} =$ heights measured at different times $t_j$.
- $x_i =$ height function of individual $i$.
Example: Berkeley Growth data

- $\{y_{ij}\}_{j=1,...,p} =$ heights measured at different times $t_j$.
- $x_i =$ height function of individual $i$.
Motivations of our study

- Most of previous works do not consider that FD belong to a RKHS. → We want to investigate kernel methods for FDC. E.g. of related work [?, ?].

- Most of previous works only use one representation \( x_i \) or \( Dx_i \) the derivative functions. → We want to investigate if information fusion can leverage the functional nature of the data by considering Sobolev spaces \( W_{1,2}([0, T]) \). E.g. of related work [?].

- Most of previous works assume that FD belong to linear spaces or subspaces. → We want to investigate further the manifold hypothesis: FD belong to low-dimensional non-linear manifold. E.g. of related work [?].

⇒ We investigate these points jointly and from an empirical viewpoint using 20 benchmarks and by using spectral clustering (SC).
Motivations of our study

- Most of previous works do not consider that FD belong to a **RKHS**.
  \[ \rightarrow \text{We want to investigate kernel methods for FDC.} \]
  \[ \text{E.g. of related work [?, ?].} \]
- Most of previous works only use one representation \( x_i \) xor \( Dx_i \) the derivative functions.
  \[ \rightarrow \text{We want to investigate if **information fusion** can leverage the} \]
  \[ \text{functional nature of the data by considering Sobolev spaces} \]
  \[ \mathcal{W}^{1,2}([0, T]). \]
  \[ \text{E.g. of related work [?].} \]
Motivations of our study

- Most of previous works do not consider that FD belong to a **RKHS**.  
  → We want to investigate kernel methods for FDC.  
    E.g. of related work [?, ?].

- Most of previous works only use one representation $x_i$ xor $Dx_i$ the derivative functions.
  → We want to investigate if **information fusion** can leverage the functional nature of the data by considering Sobolev spaces $W^{1,2}([0, T])$.  
    E.g. of related work [?].

- Most of previous works assume that FD belong to linear spaces or subspaces.
  → We want to investigate further the **manifold hypothesis** : FD belong to low-dimensional non-linear manifold.  
    E.g. of related work [?].
Motivations of our study

- Most of previous works do not consider that FD belong to a RKHS.
  → We want to investigate kernel methods for FDC.
    E.g. of related work [?, ?].
- Most of previous works only use one representation $x_i$ xor $Dx_i$ the derivative functions.
  → We want to investigate if information fusion can leverage the functional nature of the data by considering Sobolev spaces $W^{1,2}([0, T])$.
    E.g. of related work [?].
- Most of previous works assume that FD belong to linear spaces or subspaces.
  → We want to investigate further the manifold hypothesis: FD belong to low-dimensional non-linear manifold.
    E.g. of related work [?].
⇒ We investigate these points jointly and from an empirical viewpoint using 20 benchmarks and by using spectral clustering (SC).
Outline

1. Background and Motivations
2. Spectral clustering
3. Experiments and discussion
Spectral clustering (SC) in a nutshell

- Methods developed in the ML community since the early 2000’s.
- Capture the intrinsic geometry of the data.
- Similarity, neighbor end Laplacian graphs are important concepts.
- Methodology: use the spectral decomposition of the Laplacian matrix as an embedding of the graph nodes in an Euclidean space then partition the nodes using $k$-means.
- Motivations: the eigenvalues and eigenvectors of the Laplacian encode information about the connected components (and more generally clusters) of the graph, they also provide solutions to (relaxed) graph cuts problems.
- See for e.g. [?] for an introduction.
Similarity and Neighbor graphs

- Similarities between objects as a **weighted undirected graph**
  \[ G = (\mathcal{V}, \mathcal{E}) : \]
  - \( \mathcal{V} = \{x_1, \ldots, x_n\} \) is the set of nodes: objects to cluster.
  - \( \mathcal{E} \) is the set of edges: pairs of objects that are similar to each other.
- Edges are weighted: if \((x_i, x_j) \in \mathcal{E}\) then \(K(x_i, x_j) > 0\) is the measure of the similarity.
- \( G \) is represented by a **weighted adjacency matrix** denoted \( \mathbf{W} = (w_{ij})_{i,j=1,...,n} \) with:
  \[
  w_{ij} = \begin{cases} 
  K(x_i, x_j) & \text{if } (x_i, x_j) \in \mathcal{E} \\
  0 & \text{else}
  \end{cases}
  \]
Similarity and Neighbor graphs

- Similarities between objects as a **weighted undirected graph** $G = (\mathcal{V}, \mathcal{E})$:
  - $\mathcal{V} = \{x_1, \ldots, x_n\}$ is the set of nodes: objects to cluster.
  - $\mathcal{E}$ is the set of edges: pairs of objects that are similar to each other.

- Edges are weighted: if $(x_i, x_j) \in \mathcal{E}$ then $K(x_i, x_j) > 0$ is the **measure of the similarity**.

- $G$ is represented by a **weighted adjacency matrix** denoted $W = (w_{ij})_{i,j=1,...,n}$ with:

  $$w_{ij} = \begin{cases} 
  K(x_i, x_j) & \text{if } (x_i, x_j) \in \mathcal{E} \\
  0 & \text{else}
  \end{cases}$$

- $K$ is a **kernel function**: objects belong to an RKHS.
Similarities between objects as a **weighted undirected graph** 

\[ G = (\mathcal{V}, \mathcal{E}) : \]

- \( \mathcal{V} = \{x_1, \ldots, x_n\} \) is the set of nodes: objects to cluster.
- \( \mathcal{E} \) is the set of edges: pairs of objects that are similar to each other.

Edges are weighted: if \((x_i, x_j) \in \mathcal{E}\) then \(K(x_i, x_j) > 0\) is the **measure of the similarity**.

- \(G\) is represented by a **weighted adjacency matrix** denoted 
\[ W = (w_{ij})_{i,j=1,...,n} \] with:

\[ w_{ij} = \begin{cases} 
K(x_i, x_j) & \text{if } (x_i, x_j) \in \mathcal{E} \\
0 & \text{else} 
\end{cases} \]

- \(K\) is a **kernel function**: objects belong to an RKHS.
- We can sparsify \(W\) and have a \(k\) **nearest neighbor graph** in order to strengthen the manifold hypothesis.
Laplacian matrix and its normalization

- Let $D = (d_{ij})_{i,j=1,...,n}$ be the **degree matrix** defined by:

  $$d_{ij} = \begin{cases} d_i & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

  with $d_i = \sum_{j=1}^{n} w_{ij}, \forall i = 1, \ldots, n$.

- The **Laplacian matrix** of $G$ denoted $L$ is given by:

  $$L = D - W$$

- Its **(symmetric) normalization** denoted $L_{sym}$ is defined by:

  $$L_{sym} = D^{-1/2} LD^{-1/2} = I - D^{-1/2} WD^{-1/2}$$

  with $I$ the identity matrix of order $n$. 
Properties of the normalized Laplacian matrix

Property.

- $L_{\text{sym}}$ can be viewed as a quadratic form (that we aim at minimizing):
  \[
  f^T L_{\text{sym}} f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left( \frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2, \forall f \in \mathbb{R}^n
  \]

- $L_{\text{sym}}$ is symmetric and psd:
  \[
  0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n
  \]

- The multiplicity order $k$ of the null eigenvalue is the number of connected components of $G$. Let denote the latter subset of nodes as $C_1, \ldots, C_k$. The eigen subspace associated to $\lambda_1$ is spanned by $D^{1/2}1_{C_1}, \ldots, D^{1/2}1_{C_k}$ where $1_{C_i}$ is the assignment vector of $C_i$. 

Illustration with a disconnected graph

\[ V = \{ x_1, x_2, x_3, x_4, x_5 \} \]

\[ E = \{ (x_1, x_2), (x_2, x_3), (x_4, x_5) \} \]
Illustration with a disconnected graph

\( V = \{x_1, x_2, x_3, x_4, x_5\} \)

\( E = \{(x_1, x_2), (x_2, x_3), (x_4, x_5)\} \)

\[ W = \begin{pmatrix} 0 & 2 & 0 & 0 & 0 \\ 2 & 0 & 3 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 2 & 0 \end{pmatrix} \]

\[ D = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix} \]

\[ L = \begin{pmatrix} 2 & -2 & 0 & 0 & 0 \\ -2 & 5 & -3 & 0 & 0 \\ 0 & -3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & -2 \\ 0 & 0 & 0 & -2 & 2 \end{pmatrix} \]

\[ L_{\text{sym}} = \begin{pmatrix} 1 & -0.63 & 0 & 0 & 0 \\ -0.63 & 1 & -0.77 & 0 & 0 \\ 0 & -0.77 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \]

Spectra of \( L_{\text{sym}} \):
\( \{2, 2, 1, 0, 0\} \)

\[ D^{1/2} C_1 f_1 = \begin{pmatrix} 1.41 \\ 2.24 \\ 1.73 \\ 0 \\ 0 \end{pmatrix} \]

\[ D^{1/2} C_2 f_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1.41 \\ 1.41 \end{pmatrix} \]
Illustration with a disconnected graph

\[ V = \{ x_1, x_2, x_3, x_4, x_5 \} \]

\[ E = \{ \{ x_1, x_2 \}, \{ x_2, x_3 \}, \{ x_4, x_5 \} \} \]

\[ W = \begin{pmatrix}
0 & 2 & 0 & 0 & 0 \\
2 & 0 & 3 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 \\
0 & 0 & 0 & 2 & 0 \\
\end{pmatrix} \]

\[ D = \begin{pmatrix}
2 & 0 & 0 & 0 & 0 \\
0 & 5 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 2 \\
\end{pmatrix} \]

\[ L = \begin{pmatrix}
2 & -2 & 0 & 0 & 0 \\
-2 & 5 & -3 & 0 & 0 \\
0 & -3 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & -2 \\
0 & 0 & 0 & -2 & 2 \\
\end{pmatrix} \]

\[ L_{sym} = \begin{pmatrix}
1 & -0.63 & 0 & 0 & 0 \\
-0.63 & 1 & -0.63 & 0 & 0 \\
0 & -0.63 & 1 & -0.63 & 0 \\
0 & 0 & -0.63 & 1 & -1 \\
0 & 0 & 0 & 1 & -1 \\
\end{pmatrix} \]
Illustration with a disconnected graph

\[ \mathcal{V} = \{x_1, x_2, x_3, x_4, x_5\} \]

\[ \mathcal{E} = \{(x_1, x_2), (x_2, x_3), (x_4, x_5)\} \]

\[ W = \begin{pmatrix}
0 & 2 & 0 & 0 & 0 \\
2 & 0 & 3 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 \\
0 & 0 & 0 & 2 & 0 \\
\end{pmatrix} \]

\[ D = \begin{pmatrix}
2 & 0 & 0 & 0 & 0 \\
0 & 5 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 2 \\
\end{pmatrix} \]
Illustration with a disconnected graph

\[ \mathcal{V} = \{ x_1, x_2, x_3, x_4, x_5 \} \]
\[ \mathcal{E} = \{(x_1, x_2), (x_2, x_3), (x_4, x_5)\} \]

\[ W = \begin{pmatrix}
0 & 2 & 0 & 0 & 0 \\
2 & 0 & 3 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 2 & 0
\end{pmatrix} \quad \rightarrow \quad D = \begin{pmatrix}
2 & 0 & 0 & 0 & 0 \\
0 & 5 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 2
\end{pmatrix} \]
Illustration with a disconnected graph

\( \mathcal{V} = \{x_1, x_2, x_3, x_4, x_5\} \)

\[ \mathcal{E} = \{(x_1, x_2), (x_2, x_3), (x_4, x_5)\} \]

\[ \mathbf{W} = \begin{pmatrix}
0 & 2 & 0 & 0 & 0 \\
2 & 0 & 3 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 \\
0 & 0 & 0 & 2 & 0
\end{pmatrix} \]

\[ \mathbf{D} = \begin{pmatrix}
2 & 0 & 0 & 0 & 0 \\
0 & 5 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 2
\end{pmatrix} \]

\[ \mathbf{L} = \begin{pmatrix}
2 & -2 & 0 & 0 & 0 \\
-2 & 5 & -3 & 0 & 0 \\
0 & -3 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & -2 \\
0 & 0 & 0 & -2 & 2
\end{pmatrix} \]

\[ \mathbf{L}_{sym} = \begin{pmatrix}
1 & -0.63 & 0 & 0 & 0 \\
-0.63 & 1 & -0.77 & 0 & 0 \\
0 & -0.77 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 \\
0 & 0 & 0 & -1 & 1
\end{pmatrix} \]
Illustration with a disconnected graph

\[ \mathcal{V} = \{x_1, x_2, x_3, x_4, x_5\} \]
\[ \mathcal{E} = \{(x_1, x_2), (x_2, x_3), (x_4, x_5)\} \]

\[ \rightarrow W = \begin{pmatrix}
0 & 2 & 0 & 0 & 0 \\
2 & 0 & 3 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 \\
0 & 0 & 0 & 2 & 0
\end{pmatrix} \]
\[ \rightarrow D = \begin{pmatrix}
2 & 0 & 0 & 0 & 0 \\
0 & 5 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 2
\end{pmatrix} \]

\[ \rightarrow L = \begin{pmatrix}
2 & -2 & 0 & 0 & 0 \\
-2 & 5 & -3 & 0 & 0 \\
0 & -3 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & -2 \\
0 & 0 & 0 & -2 & 2
\end{pmatrix} \]
\[ \rightarrow L_{sym} = \begin{pmatrix}
1 & -0.63 & 0 & 0 & 0 \\
-0.63 & 1 & -0.77 & 0 & 0 \\
0 & -0.77 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 \\
0 & 0 & 0 & -1 & 1
\end{pmatrix} \]

→ Spectra of \( L_{sym} \): \{2,2,1,0,0\}
Illustration with a disconnected graph

\[ \mathcal{V} = \{x_1, x_2, x_3, x_4, x_5\} \]
\[ \mathcal{E} = \{(x_1, x_2), (x_2, x_3), (x_4, x_5)\} \]

\[ \rightarrow W = \begin{pmatrix}
0 & 2 & 0 & 0 & 0 \\
2 & 0 & 3 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 \\
0 & 0 & 0 & 2 & 0
\end{pmatrix} \]

\[ \rightarrow D = \begin{pmatrix}
2 & 0 & 0 & 0 & 0 \\
0 & 5 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 2
\end{pmatrix} \]

\[ \rightarrow L = \begin{pmatrix}
2 & -2 & 0 & 0 & 0 \\
-2 & 5 & -3 & 0 & 0 \\
0 & -3 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & -2 \\
0 & 0 & 0 & -2 & 2
\end{pmatrix} \]

\[ \rightarrow L_{sym} = \begin{pmatrix}
1 & -0.63 & 0 & 0 & 0 \\
-0.63 & 1 & -0.77 & 0 & 0 \\
0 & -0.77 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 \\
0 & 0 & 0 & -1 & 1
\end{pmatrix} \]

\[ \rightarrow \text{Spectra of } L_{sym} : \{2, 2, 1, 0, 0\} \]

\[ \rightarrow D^{1/2}1_{C_1} = \begin{pmatrix}
1.41 \\
2.24 \\
1.73 \\
0 \\
0
\end{pmatrix} \quad \text{and} \quad D^{1/2}1_{C_2} = \begin{pmatrix}
0 \\
0 \\
1.41 \\
1.41
\end{pmatrix} \]
Outline

1. Background and Motivations

2. Spectral clustering

3. Experiments and discussion
### Workflow

1. **Smoothing**: from \( \{y_{ij}\}_{i,j} \) to \( \{x_i\}_i \):
   - Basis functions are cubic bspline \( \{\phi_k\}_{k=1,\ldots,q} \) with \( q = 4 + p \):
     \[
     x_i(t) = \mathbf{c}_i^\top \phi(t) = \sum_{k=1}^{q} c_{ik} \phi_k(t)
     \]
     where \( \mathbf{c}_i = (c_{i1} \ldots c_{iq})^\top \) and \( \phi(t) = (\phi_1(t) \ldots \phi_q(t))^\top \in \mathbb{R}^q \).
   - We find \( \mathbf{c}_i \) as follows:
     \[
     \mathbf{c}_i = \arg \min_{\mathbf{c} \in \mathbb{R}^q} \sum_{j=1}^{p} (y_{ij} - x_i(t_j))^2 + \lambda \int_0^T D^2 x_i(t) dt
     \]
     where \( D \) is the differential operator and \( \lambda \) is the smoothing coefficient selected in \( \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\} \) wrt the GCV criterion.

2. Center the \( \{x_i\}_i \) and compute derivatives \( \{Dx_i\}_i \).

3. Compute the Gram matrix \( S \) wrt a given kernel function.

4. Perform clustering procedures.

5. Evaluate clustering outputs and compare the results.
Workflow

1. **Smoothing**: from \( \{y_{ij}\}_{i,j} \) to \( \{x_i\}_i \):
   - Basis functions are cubic bspline \( \{\phi_k\}_{k=1,...,q} \) with \( q = 4 + p \):
     \[
     x_i(t) = c_i^\top \phi(t) = \sum_{k=1}^{q} c_{ik} \phi_k(t)
     \]
     where \( c_i = (c_{i1} \ldots c_{iq})^\top \) and \( \phi(t) = (\phi_1(t) \ldots \phi_q(t))^\top \in \mathbb{R}^q \).
   - We find \( c_i \) as follows:
     \[
     c_i = \arg \min_{c \in \mathbb{R}^q} \sum_{j=1}^{p} (y_{ij} - x_i(t_j))^2 + \lambda \int_0^T D^2 x_i(t) dt
     \]
     where \( D \) is the differential operator and \( \lambda \) is the smoothing coefficient selected in \( \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\} \) wrt the GCV criterion.

2. **Center** the \( \{x_i\}_i \) and compute **derivatives** \( \{Dx_i\}_i \).
Workflow

1. **Smoothing**: from \( \{y_{ij}\}_{i,j} \) to \( \{x_i\}_i \):
   - Basis functions are cubic bspline \( \{\phi_k\}_{k=1,...,q} \) with \( q = 4 + p \):
     \[
     x_i(t) = \mathbf{c}_i^\top \phi(t) = \sum_{k=1}^{q} c_{ik} \phi_k(t)
     \]
     where \( \mathbf{c}_i = (c_{i1} \ldots c_{iq})^\top \) and \( \phi(t) = (\phi_1(t) \ldots \phi_q(t))^\top \in \mathbb{R}^q \).
   - We find \( \mathbf{c}_i \) as follows:
     \[
     \mathbf{c}_i = \arg \min_{\mathbf{c} \in \mathbb{R}^q} \sum_{j=1}^{p} (y_{ij} - x_i(t_j))^2 + \lambda \int_0^T D^2 x_i(t) dt
     \]
     where \( D \) is the differential operator and \( \lambda \) is the smoothing coefficient selected in \( \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\} \) wrt the GCV criterion.

2. **Center** the \( \{x_i\}_i \) and compute derivatives \( \{Dx_i\}_i \).

3. Compute the **Gram matrix** \( \mathbf{S} \) wrt a given kernel function.
Workflow

1. **Smoothing**: from \( \{y_{ij}\}_{i,j} \) to \( \{x_i\}_i \):
   - Basis functions are cubic b spline \( \{\phi_k\}_{k=1,\ldots,q} \) with \( q = 4 + p \):
     \[
     x_i(t) = c_i^\top \phi(t) = \sum_{k=1}^{q} c_{ik} \phi_k(t)
     \]
     where \( c_i = (c_{i1} \ldots c_{iq})^\top \) and \( \phi(t) = (\phi_1(t) \ldots \phi_q(t))^\top \in \mathbb{R}^q \).
   - We find \( c_i \) as follows:
     \[
     c_i = \arg \min_{c \in \mathbb{R}^q} \sum_{j=1}^{p} (y_{ij} - x_i(t_j))^2 + \lambda \int_0^T D^2 x_i(t) \, dt
     \]
     where \( D \) is the differential operator and \( \lambda \) is the smoothing coefficient selected in \( \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\} \) wrt the GCV criterion.

2. **Center** the \( \{x_i\}_i \) and compute derivatives \( \{Dx_i\}_i \).
3. Compute the **Gram matrix** \( S \) wrt a given kernel function.
4. Perform **clustering procedures**.
Experiments and discussion

Workflow

1. **Smoothing** : from \( \{y_{ij}\}_{i,j} \) to \( \{x_i\}_i \):
   - Basis functions are cubic bspline \( \{\phi_k\}_{k=1,...,q} \) with \( q = 4 + p \):

   \[
   x_i(t) = c_i^\top \phi(t) = \sum_{k=1}^{q} c_{ik} \phi_k(t)
   \]

   where \( c_i = (c_{i1} \ldots c_{iq})^\top \) and \( \phi(t) = (\phi_1(t) \ldots \phi_q(t))^\top \in \mathbb{R}^q \).
   - We find \( c_i \) as follows:

   \[
   c_i = \arg \min_{c \in \mathbb{R}^q} \sum_{j=1}^{p} (y_{ij} - x_i(t_j))^2 + \lambda \int_0^T D^2 x_i(t) dt
   \]

   where \( D \) is the differential operator and \( \lambda \) is the smoothing coefficient selected in \( \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\} \) wrt the GCV criterion.

2. **Center** the \( \{x_i\}_i \) and compute derivatives \( \{Dx_i\}_i \).

3. Compute the **Gram matrix** \( S \) wrt a given kernel function.

4. Perform **clustering procedures**.

5. **Evaluate** clustering outputs and **compare** the results.
Kernel/Representation/Sparsification

- FD are centered: \( \sum_{i=1}^{n} x_i(t) = 0, \forall t \in [0, T] \).

Different Hilbert spaces:
- \( x_i \in L^2([0, T]) \), e.g. \( KL(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} \).
- \( dx_i \in L^2([0, T]) \), e.g. \( KL(x_i, x_j) = \langle dx_i, dx_j \rangle_{L^2} \).
- \( x_i \in W^{1,2}([0, T]) \), e.g. \( KL(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} + \langle dx_i, dx_j \rangle_{L^2} \).

Different kernel functions (RKHS):
- Linear kernel: \( KL(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} = \int_0^T x_i(t)x_j(t)dt \).
- Gaussian Kernel: \( KL(x_i, x_j) = \exp\left(-\frac{\|x_i-x_j\|^2}{\sigma_i\sigma_j}\right) \).

Different sparsifications:
- "Connected" graph: \( w_{ij} = \max(KL(x_i, x_j), 0) \).
- \( k \) nearest-neighbor graph (with \( k = 7 \)).

⇒ Main questions:
- Does basis expansion and RKHS help?
- Does "fusing" both \( x_i \) and \( dx_i \) and work in a Sobolev space help?
- Does sparsification (that emphasizes the manifold hypothesis) help?
Experiments and discussion

Kernel/Representation/Sparsification

- FD are centered: $\sum_{i=1}^{n} x_i(t) = 0, \forall t \in [0, T]$.
- Different **Hilbert spaces**:
  1. $x_i \in L^2([0, T]),$ e.g. $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2}$.
  2. $Dx_i \in L^2([0, T]),$ e.g. $K_l(x_i, x_j) = \langle Dx_i, Dx_j \rangle_{L^2}$.
  3. $x_i \in W^{1,2}([0, T]),$ e.g. $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} + \langle Dx_i, Dx_j \rangle_{L^2}$.

Different kernel functions (RKHS):
- **Linear kernel**:
  $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} = \int_{0}^{T} x_i(t)x_j(t) \, dt$.
- **Gaussian Kernel**: $K_g(x_i, x_j) = \exp(-\|x_i-x_j\|_2^2/\sigma^2)$. 

Different sparsifications:
- "Connected" graph: $w_{ij} = \max(K_l(x_i, x_j), 0)$.
- $k$ nearest-neighbor graph (with $k=7$).

$\Rightarrow$ Main questions:
- Does basis expansion and RKHS help?
- Does "fusing" both $x_i$ and $Dx_i$ and work in a Sobolev space help?
- Does sparsification (that emphasizes the manifold hypothesis) help?
FD are centered: \( \sum_{i=1}^{n} x_i(t) = 0, \forall t \in [0, T] \).

Different **Hilbert spaces**:

- **00** \( x_i \in L^2([0, T]) \), e.g. \( K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} \).
- **11** \( Dx_i \in L^2([0, T]) \), e.g. \( K_l(x_i, x_j) = \langle Dx_i, Dx_j \rangle_{L^2} \).
- **01** \( x_i \in W^{1,2}([0, T]) \), e.g. \( K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} + \langle Dx_i, Dx_j \rangle_{L^2} \).

Different **kernel functions** (**RKHS**):

- Linear kernel: \( K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} = \int_{0}^{T} x_i(t)x_j(t)dt \)
- Gaussian Kernel [?]: \( K_g(x_i, x_j) = \exp \left( -\frac{\|x_i-x_j\|_{L^2}^2}{\sigma_i\sigma_j} \right) \)
Kernel/Representation/Sparsification

- FD are centered: $\sum_{i=1}^{n} x_i(t) = 0, \forall t \in [0, T]$.
- Different **Hilbert spaces**:
  - 00 $x_i \in L^2([0, T])$, e.g. $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2}$.
  - 11 $Dx_i \in L^2([0, T])$, e.g. $K_l(x_i, x_j) = \langle Dx_i, Dx_j \rangle_{L^2}$.
  - 01 $x_i \in W^{1,2}([0, T])$, e.g. $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} + \langle Dx_i, Dx_j \rangle_{L^2}$.
- Different **kernel functions** (RKHS):
  - Linear kernel: $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} = \int_0^T x_i(t)x_j(t)dt$
  - Gaussian Kernel [?]: $K_g(x_i, x_j) = \exp\left(-\frac{\|x_i-x_j\|_{L^2}^2}{\sigma_i\sigma_j}\right)$
- Different **sparsifications**:
  - 0 “Connected” graph: $w_{ij} = \max(K(x_i, x_j), 0)$.
  - 1 $k$ nearest-neighbor graph (with $k=7$).
Kernel/Representation/Sparsification

- **FD** are centered: $\sum_{i=1}^{n} x_i(t) = 0, \forall t \in [0, T]$.
- Different **Hilbert spaces**:
  00 $x_i \in L^2([0, T]),$ e.g. $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2}$.
  11 $Dx_i \in L^2([0, T]),$ e.g. $K_l(x_i, x_j) = \langle Dx_i, Dx_j \rangle_{L^2}$.
  01 $x_i \in W^{1, 2}([0, T]),$ e.g. $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} + \langle Dx_i, Dx_j \rangle_{L^2}$.
- Different **kernel functions** (RKHS):
  - Linear kernel: $K_l(x_i, x_j) = \langle x_i, x_j \rangle_{L^2} = \int_0^T x_i(t)x_j(t)dt$
  - Gaussian Kernel [?]: $K_g(x_i, x_j) = \exp\left(-\frac{\|x_i-x_j\|_{L^2}^2}{\sigma_i \sigma_j}\right)$
- Different **sparsifications**:
  0 “Connected” graph: $w_{ij} = \max(K(x_i, x_j), 0)$.
  1 $k$ nearest-neighbor graph (with $k=7$).

⇒ **Main questions**:
- Does basis expansion and RKHS help?
- Does “fusing” both $x_i$ and $Dx_i$ and work in a Sobolev space help?
- Does sparsification (that emphasizes the manifold hypothesis) help?
Clustering procedures

- We test the different kernel/representation/sparsification using the two following clustering procedures. $S$ is the Gram matrix.
Cluster procedures

- We test the different kernel/representation/sparsification using the two following clustering procedures. \( S \) is the Gram matrix.
  - **Kernel \( k \)-means (\( K_{km} \))**:
    - Spectral decomposition of \( S \).
    - Euclidean embedding: \( F = (f_1 \ldots f_l) \) (all eigenvectors associated to strictly positive eigenvalues).
    - Apply \( k \)-means to \( F \).
We test the different kernel/representation/sparsification using the two following clustering procedures. $S$ is the Gram matrix.

- **Kernel $k$-means ($K_{km}$):**
  - Spectral decomposition of $S$.
  - Euclidean embedding: $F = (f_1 \ldots f_l)$ (all eigenvectors associated to strictly positive eigenvalues).
  - Apply $k$-means to $F$.

- **Spectral clustering ($SC_{km}$):**
  - From $S$, determine $W$ (with/without sparsification) and $L_{sym}$.
  - Spectral decomposition of $L_{sym}$.
  - Euclidean embedding: $F = (f_1 \ldots f_k)$ ($k$ first eigenvectors associated to the lowest eigenvalues).
  - Normalize rows of $F$ to have unit norms.
  - Apply $k$-means to $F$. 
Experiments and discussion

Clustering procedures

- We test the different kernel/representation/sparsification using the two following clustering procedures. \(S\) is the Gram matrix.
  - **Kernel \(k\)-means (\(K_{km}\))**: 
    - Spectral decomposition of \(S\).
    - Euclidean embedding: \(F = (f_1 \ldots f_l)\) (all eigenvectors associated to strictly positive eigenvalues).
    - Apply \(k\)-means to \(F\).
  - **Spectral clustering (\(SC_{km}\))**: 
    - From \(S\), determine \(W\) (with/without sparsification) and \(L_{sym}\).
    - Spectral decomposition of \(L_{sym}\).
    - Euclidean embedding: \(F = (f_1 \ldots f_k)\) (\(k\) first eigenvectors associated to the lowest eigenvalues).
    - Normalize rows of \(F\) to have unit norms.
    - Apply \(k\)-means to \(F\).
  - **Baseline**: kernel \(k\)-means with linear kernel \(K_l(x_i, x_j) = \langle x_i, x_j \rangle_{\mathbb{L}^2}\).
List of the 18 clustering models

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Representation</th>
<th>Kernel</th>
<th>Clustering proc.</th>
<th>Sparsif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>00_linear_K_km</td>
<td>( x_i \in \mathbb{L}^2([0, T]) )</td>
<td>Linear</td>
<td>Ker. k-means</td>
<td></td>
</tr>
</tbody>
</table>
### Experiments and discussion

List of the 18 clustering models

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Representation</th>
<th>Kernel</th>
<th>Clustering proc.</th>
<th>Sparsif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>00_linear_K_km</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>11_linear_K_km</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>01_linear_K_km</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Linear</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
</tbody>
</table>
# List of the 18 clustering models

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Representation</th>
<th>Kernel</th>
<th>Clustering proc.</th>
<th>Sparsif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>00_gaussian_K_km</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>11_gaussian_K_km</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>01_gaussian_K_km</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Gaussian</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
</tbody>
</table>
List of the 18 clustering models

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Representation</th>
<th>Kernel</th>
<th>Clustering proc.</th>
<th>Sparsif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>00_linear_K_km</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>00_gaussian_K_km</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>11_linear_K_km</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>11_gaussian_K_km</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>01_linear_K_km</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Linear</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
<tr>
<td>01_gaussian_K_km</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Gaussian</td>
<td>Ker. $k$-means</td>
<td></td>
</tr>
</tbody>
</table>
List of the 18 clustering models

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Representation</th>
<th>Kernel</th>
<th>Clustering proc.</th>
<th>Sparsif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>00_linear_K_km_</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>00_gaussian_K_km_</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>11_linear_K_km_</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>11_gaussian_K_km_</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>01_linear_K_km_</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Linear</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>01_gaussian_K_km_</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Gaussian</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>00_linear_SC_km_0</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust.</td>
<td>Connected</td>
</tr>
<tr>
<td>00_gaussian_SC_km_0</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust.</td>
<td>Connected</td>
</tr>
<tr>
<td>11_linear_SC_km_0</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust.</td>
<td>Connected</td>
</tr>
<tr>
<td>11_gaussian_SC_km_0</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust.</td>
<td>Connected</td>
</tr>
<tr>
<td>01_linear_SC_km_0</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust.</td>
<td>Connected</td>
</tr>
<tr>
<td>01_gaussian_SC_km_0</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust.</td>
<td>Connected</td>
</tr>
</tbody>
</table>
## List of the 18 clustering models

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Representation</th>
<th>Kernel</th>
<th>Clustering proc.</th>
<th>Sparsif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>00_linear_K_km</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>00_gaussian_K_km</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>11_linear_K_km</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>11_gaussian_K_km</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>01_linear_K_km</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Linear</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>01_gaussian_K_km</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Gaussian</td>
<td>Ker. k-means</td>
<td></td>
</tr>
<tr>
<td>00_linear_SC_km_0</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust.  Connected</td>
<td></td>
</tr>
<tr>
<td>00_gaussian_SC_km_0</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust. Connected</td>
<td></td>
</tr>
<tr>
<td>11_linear_SC_km_0</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust. Connected</td>
<td></td>
</tr>
<tr>
<td>11_gaussian_SC_km_0</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust. Connected</td>
<td></td>
</tr>
<tr>
<td>01_linear_SC_km_0</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust. Connected</td>
<td></td>
</tr>
<tr>
<td>01_gaussian_SC_km_0</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust. Connected</td>
<td></td>
</tr>
<tr>
<td>00_linear_SC_km_1</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust. 7 near. neig.</td>
<td></td>
</tr>
<tr>
<td>00_gaussian_SC_km_1</td>
<td>$x_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust. 7 near. neig.</td>
<td></td>
</tr>
<tr>
<td>11_linear_SC_km_1</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust. 7 near. neig.</td>
<td></td>
</tr>
<tr>
<td>11_gaussian_SC_km_1</td>
<td>$Dx_i \in \mathbb{L}^2([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust. 7 near. neig.</td>
<td></td>
</tr>
<tr>
<td>01_linear_SC_km_1</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Linear</td>
<td>Spectral Clust. 7 near. neig.</td>
<td></td>
</tr>
<tr>
<td>01_gaussian_SC_km_1</td>
<td>$x_i \in \mathbb{W}^{1,2}([0, T])$</td>
<td>Gaussian</td>
<td>Spectral Clust. 7 near. neig.</td>
<td></td>
</tr>
</tbody>
</table>
## 20 datasets from fda, fda.usc and UCR_TS_Archive_2015

<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th>Name</th>
<th>Nb of FD</th>
<th>Nb of Class</th>
<th>Nb of obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>fda</td>
<td>Growth</td>
<td>93</td>
<td>2</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>fda.usc</td>
<td>poblenu</td>
<td>115</td>
<td>2</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>fda.usc</td>
<td>tecator</td>
<td>215</td>
<td>2</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>fda.usc</td>
<td>phoneme</td>
<td>250</td>
<td>5</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Spectro</td>
<td>Beef</td>
<td>30</td>
<td>5</td>
<td>470</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Simulated</td>
<td>CBF</td>
<td>30</td>
<td>3</td>
<td>128</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Spectro</td>
<td>Coffee</td>
<td>28</td>
<td>2</td>
<td>286</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>ECG</td>
<td>ECG200</td>
<td>100</td>
<td>2</td>
<td>96</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Image</td>
<td>FaceFour</td>
<td>24</td>
<td>4</td>
<td>350</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Image</td>
<td>Fish</td>
<td>175</td>
<td>7</td>
<td>463</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Motion</td>
<td>GunPoint</td>
<td>50</td>
<td>2</td>
<td>150</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Sensor</td>
<td>Lightning2</td>
<td>60</td>
<td>2</td>
<td>637</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Sensor</td>
<td>Lightning7</td>
<td>70</td>
<td>7</td>
<td>319</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Image</td>
<td>MedicalImages</td>
<td>381</td>
<td>10</td>
<td>99</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Spectro</td>
<td>OliveOil</td>
<td>30</td>
<td>4</td>
<td>570</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Image</td>
<td>OSULeaf</td>
<td>200</td>
<td>6</td>
<td>427</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Image</td>
<td>SwedishLeaf</td>
<td>500</td>
<td>15</td>
<td>128</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Image</td>
<td>Symbols</td>
<td>25</td>
<td>6</td>
<td>398</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Sensor</td>
<td>Trace</td>
<td>100</td>
<td>4</td>
<td>275</td>
</tr>
<tr>
<td>UCR_TS</td>
<td>Simulated</td>
<td>TwoPatterns</td>
<td>1000</td>
<td>4</td>
<td>128</td>
</tr>
</tbody>
</table>
Clustering assessment and comparison

- Clustering models assessment:
  - **External validation**: for each dataset we have the ground truth.
  - Compare a clustering output against the ground truth using the **Normalized Mutual Information (NMI)** criterion. This measure is between 0 and 1 and the bigger the better.

---

1. *i* beats *j* for a given dataset, if NMI of *i* > NMI of *j*
Clustering assessment and comparison

- Clustering models assessment :
  - **External validation**: for each dataset we have the ground truth.
  - Compare a clustering output against the ground truth using the **Normalized Mutual Information (NMI)** criterion. This measure is between 0 and 1 and the bigger the better.

- Comparing the 18 clustering models :
  - For each pair of clustering models \((i, j)\), we count the **nb of times** \(i\) **beats** \(j\) among the 20 datasets (each dataset is seen as a “match”).
  - For an overall ranking of the clustering models, we use **Borda’s voting rule**: we rank according to the total **nb of wins**. Each clustering model “plays” in total \(20 \times 17 = 340\) “matches”.

---

1. \(i\) beats \(j\) for a given dataset, if NMI of \(i\) > NMI of \(j\)
Examples of results: Growth data

- **00_linear**
  - nmi: 0.0 0.2 0.4 0.6 0.8 1.0
  - K_km, SC_km_0, SC_km_1

- **11_linear**
  - nmi: 0.0 0.2 0.4 0.6 0.8 1.0
  - K_km, SC_km_0, SC_km_1

- **01_linear**
  - nmi: 0.0 0.2 0.4 0.6 0.8 1.0
  - K_km, SC_km_0, SC_km_1

- **00_gaussian**
  - nmi: 0.0 0.2 0.4 0.6 0.8 1.0
  - K_km, SC_km_0, SC_km_1

- **11_gaussian**
  - nmi: 0.0 0.2 0.4 0.6 0.8 1.0
  - K_km, SC_km_0, SC_km_1

- **01_gaussian**
  - nmi: 0.0 0.2 0.4 0.6 0.8 1.0
  - K_km, SC_km_0, SC_km_1

J. Ah-Pine, A-F. Yao
A study of manifold hypothesis for FD
CMStatistics 2019
Examples of results: SwedishLeaf data
Examples of results: Fish data
Examples of results: Tecator data
### Overall results: Borda’s ranking

<table>
<thead>
<tr>
<th>Rank</th>
<th>Clustering model</th>
<th>Nb of wins</th>
<th>Nb of losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01_gaussian_SC_km_1</td>
<td>206</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>11_gaussian_K_km_</td>
<td>195</td>
<td>113</td>
</tr>
<tr>
<td>3</td>
<td>00_gaussian_K_km_</td>
<td>182</td>
<td>119</td>
</tr>
<tr>
<td>4</td>
<td>00_gaussian_SC_km_1</td>
<td>179</td>
<td>125</td>
</tr>
<tr>
<td>5</td>
<td>01_gaussian_SC_km_0</td>
<td>175</td>
<td>131</td>
</tr>
<tr>
<td>6</td>
<td>11_gaussian_SC_km_1</td>
<td>174</td>
<td>136</td>
</tr>
<tr>
<td>7</td>
<td>01_gaussian_K_km_</td>
<td>170</td>
<td>134</td>
</tr>
<tr>
<td>8</td>
<td>00_linear_SC_km_0</td>
<td>151</td>
<td>143</td>
</tr>
<tr>
<td>9</td>
<td>11_gaussian_SC_km_0</td>
<td>151</td>
<td>159</td>
</tr>
<tr>
<td>10</td>
<td>00_linear_SC_km_1</td>
<td>150</td>
<td>159</td>
</tr>
<tr>
<td>11</td>
<td>01_linear_SC_km_0</td>
<td>148</td>
<td>146</td>
</tr>
<tr>
<td>12</td>
<td>01_linear_SC_km_1</td>
<td>146</td>
<td>157</td>
</tr>
<tr>
<td>13</td>
<td>00_gaussian_SC_km_0</td>
<td>131</td>
<td>167</td>
</tr>
<tr>
<td>14</td>
<td>11_linear_SC_km_0</td>
<td>127</td>
<td>177</td>
</tr>
<tr>
<td>15</td>
<td>11_linear_K_km_</td>
<td>117</td>
<td>193</td>
</tr>
<tr>
<td>16</td>
<td>01_linear_K_km_</td>
<td>117</td>
<td>190</td>
</tr>
<tr>
<td><em>17</em></td>
<td>00_linear_K_km_</td>
<td>115</td>
<td>193</td>
</tr>
<tr>
<td>18</td>
<td>11_linear_SC_km_1</td>
<td>104</td>
<td>202</td>
</tr>
</tbody>
</table>
## Overall results: Borda’s ranking

<table>
<thead>
<tr>
<th>Rank</th>
<th>Clustering model</th>
<th>Nb of wins</th>
<th>Nb of losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01_gaussian_SC_km_1</td>
<td>206</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>11_gaussian_K_km_</td>
<td>195</td>
<td>113</td>
</tr>
<tr>
<td>3</td>
<td>00_gaussian_K_km_</td>
<td>182</td>
<td>119</td>
</tr>
<tr>
<td>4</td>
<td>00_gaussian_SC_km_1</td>
<td>179</td>
<td>125</td>
</tr>
<tr>
<td>5</td>
<td>01_gaussian_SC_km_0</td>
<td>175</td>
<td>131</td>
</tr>
<tr>
<td>6</td>
<td>11_gaussian_SC_km_1</td>
<td>174</td>
<td>136</td>
</tr>
<tr>
<td>7</td>
<td>01_gaussian_K_km_</td>
<td>170</td>
<td>134</td>
</tr>
<tr>
<td>9</td>
<td>11_gaussian_SC_km_0</td>
<td>151</td>
<td>159</td>
</tr>
<tr>
<td>13</td>
<td>00_gaussian_SC_km_0</td>
<td>131</td>
<td>167</td>
</tr>
</tbody>
</table>
## Overall results: Borda’s ranking

<table>
<thead>
<tr>
<th>Rank</th>
<th>Clustering model</th>
<th>Nb of wins</th>
<th>Nb of losses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01_gaussian_SC_km_1</td>
<td>206</td>
<td>94</td>
</tr>
<tr>
<td>5</td>
<td>01_gaussian_SC_km_0</td>
<td>175</td>
<td>131</td>
</tr>
<tr>
<td>7</td>
<td>01_gaussian_K_km_</td>
<td>170</td>
<td>134</td>
</tr>
</tbody>
</table>
Borda’s ranking visualization
Wrap up and future work

- Given a clustering procedure, say SC\_km, we observe that:
  - Gaussian kernel gives better results than linear kernel.
Given a clustering procedure, say SC_km, we observe that:

- Gaussian kernel gives better results than linear kernel.
- Depending on the datasets $x_i$, $Dx_i$ and $(x_i, Dx_i)$ can give variable results, BUT $(x_i, Dx_i)$ is never the worst performance of the three.
Given a clustering procedure, say SC\_km, we observe that:

- Gaussian kernel gives better results than linear kernel.
- Depending on the datasets $x_i$, $Dx_i$ and $(x_i, Dx_i)$ can give variable results, BUT $(x_i, Dx_i)$ is never the worst performance of the three.
- With Gaussian kernel, $k = 7$ nearest neighbor graph sparsification always performs better than the “connected” graph sparsification. This outcome supports the manifold hypothesis.

Future work:

- $x_i$ and $Dx_i$ seems to bring complementary information BUT a “simple” fusion might degrade the overall performance.

⇒ Sparse clustering in Sobolev spaces: select discriminant features while performing the clustering.
Wrap up and future work

- Given a clustering procedure, say SC_km, we observe that:
  - Gaussian kernel gives better results than linear kernel.
  - Depending on the datasets $x_i$, $Dx_i$ and $(x_i, Dx_i)$ can give variable results, BUT $(x_i, Dx_i)$ is never the worst performance of the three.
  - With Gaussian kernel, $k = 7$ nearest neighbor graph sparsification always performs better than the “connected” graph sparsification. This outcome supports the manifold hypothesis.

- Future work:
Given a clustering procedure, say SC_km, we observe that:

- Gaussian kernel gives better results than linear kernel.
- Depending on the datasets $x_i$, $Dx_i$ and $(x_i, Dx_i)$ can give variable results, BUT $(x_i, Dx_i)$ is never the worst performance of the three.
- With Gaussian kernel, $k = 7$ nearest neighbor graph sparsification always performs better than the “connected” graph sparsification. This outcome supports the manifold hypothesis.

Future work:

- $x_i$ and $Dx_i$ seems to bring complementary information BUT a “simple” fusion might degrade the overall performance.
Given a clustering procedure, say SC\_km, we observe that:

- Gaussian kernel gives better results than linear kernel.
- Depending on the datasets $x_i$, $Dx_i$ and $(x_i, Dx_i)$ can give variable results, BUT $(x_i, Dx_i)$ is never the worst performance of the three.
- With Gaussian kernel, $k = 7$ nearest neighbor graph sparsification always performs better than the “connected” graph sparsification. This outcome supports the manifold hypothesis.

Future work:

- $x_i$ and $Dx_i$ seems to bring complementary information BUT a “simple” fusion might degrade the overall performance.

$\Rightarrow$ Sparse clustering in Sobolev spaces: select discriminant features while performing the clustering.
Thank you for your attention!
Any question or comment? :-}
Some references I

Unsupervised curve clustering using b-splines.
Scandinavian journal of statistics, 30(3) :581–595.

Infinitely divisible matrices.

Model-based clustering of time series in group-specific functional subspaces.
Advances in Data Analysis and Classification, 5(4) :281–300.

Nonlinear manifold representations for functional data.

Spectral graph theory, volume 92.
American Mathematical Soc.

Kernel k-means : spectral clustering and normalized cuts.
In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 551–556. ACM.

Nonparametric functional data analysis : theory and practice.
Some references II

Algebraic connectivity of graphs.
Czechoslovak mathematical journal, 23(2) :298–305.

Sparse clustering of functional data.

K-means algorithms for functional data.
Neurocomputing, 151 :231–245.

Functional data clustering : a survey.
Advances in Data Analysis and Classification, 8(3) :231–255.

Learning segmentation by random walks.
In NIPS, volume 14.

Representing functional data using support vector machines.
Pattern Recognition Letters, 31(6) :511–516.

On spectral clustering : Analysis and an algorithm.
In NIPS, volume 14, pages 849–856.
Some references III


