Statistical data science and some unsupervised learning problems

Faicel Chamroukhi

https://chamroukhi.com/

Workshop DSCI, 22-24 dec 2018
Outline

1. Statistics and Data Science
2. Clustering of multivariate data
3. Unsupervised learning for dimensionality reduction
4. Time series segmentation
5. Clustering of functional data
6. Unsupervised Bayesian (non-)parametric learning
7. Model-Based Co-Clustering of Multivariate Functional Data
The term “Data Science” has surged in popularity due to its increasing common use with “big data.” Data science, including Big Data, has recently attracted an enormous interest from the scientific community.
Faicel Chamroukhi
Statistical data science and some unsupervised learning problems
What does Data Science mean?

What about Statistics in the Data Science “area”?

There is not yet a consensus on what precisely constitutes Data Science.

For a review, see the report of D. Donoho (2015) : “50 years of Data Science”
Faïcel Chamroukhi

Statistical data science and some unsupervised learning problems

Introduction
Serge Abiteboul, directeur de recherche Inria, École normale supérieure de Cachan, membre de l'Académie des sciences et Patrick Flondrin, directeur de recherche CNRS, École normale supérieure de Lyon, membre de l'Académie des sciences

À la découverte des connaissances massives de la Toile
Serge Abiteboul, directeur de recherche Inria, École normale supérieure de Cachan, membre de l'Académie des sciences

Des mathématiques pour l'analyse de données massives
Stéphane Mallat, professeur à l'École normale supérieure, Paris

La découverte du cerveau grâce à l'exploration de données massives
Anastasia Ailamaki, professeure à l’École polytechnique fédérale de Lausanne

Big Data et Relaition Client : quel impact sur les industries et activités de services traditionnelles ?
François Bourdoncle, co-fondateur et CTO d’Exalead, filiale de Dassault Systèmes

Discussion générale et conclusion

Vidéos réalisées par la cellule Webocast CC-IN2P3 du CNRS
There is not yet a consensus on what precisely constitutes Data Science, but Data Science can be seen (defined?) as:

- the study of the generalizable extraction of knowledge from data.
- requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization.

---

Data Science clearly has an interdisciplinary nature and requires substantial collaborative effort.

Databases, statistics and machine learning, and distributed systems are emerging as foundational to data science.

(i) Databases : organization of data resources,
(ii) Statistics and Machine Learning : convert data into knowledge,
(iii) Distributed and Parallel Systems : computational infrastructure
Statistics play a central role in data science

- Allow to quantify the randomness component in the data
- A well-established background to deal with uncertainty (probabilistic framework) and to establish generizable methods for prediction and estimation
- Allow soft decision: e.g., confidence interval in regression and posterior probabilities in classification
- Help for understanding the underlying generative process
Statistical modeling for data science

- The observed data \((x_1, \ldots, x_n)\) where \(x_i \in \mathcal{X} \subseteq \mathbb{R}^d\) are assumed to represent samples from random variables \(X\) with unknown probability distribution \(f\).

- The main questions are i) how to define flexible and generic models for \(f\) ii) construct estimators with desirable properties to learn \(f\) from the data iii) to deal with the computational and practical issues for “complex” data.

- The area of statistical learning for the analysis of complex data.

Context and Objectives

**Context**: Large-scale data are increasingly frequent: Complex data ↦ heterogeneous, dynamical (temporal, functional), incomplete, high-dimension, and possibly massive.

**Objectives**: learn/discern useful information in an unsupervised way from raw data:

↩ Reconstruct/reveal hidden structures, i.e, (hierarchy of) groups; learn/select relevant features, etc.
Unsupervised Sparse Signal Decomposition

![Image of whale](image_url)
Unsupervised Sparse Signal Decomposition
Unsupervised Sparse Signal Decomposition
Unsupervised Sparse Signal Decomposition
1. Statistics and Data Science

2. Clustering of multivariate data

3. Unsupervised learning for dimensionality reduction

4. Time series segmentation

5. Clustering of functional data

6. Unsupervised Bayesian (non-)parametric learning

7. Model-Based Co-Clustering of Multivariate Functional Data
Clustering of multivariate data

Geyser Data

Faicel Chamroukhi

Statistical data science and some unsupervised learning problems
Clustering of multivariate data

Geyser Data clustering K–means : K–means iteration : 6

Faicel Chamroukhi
Statistical data science and some unsupervised learning problems
**K-means**

- a straightforward and widely used clustering algorithm, is one of the most important algorithms in unsupervised learning.
- Each cluster is represented by its mean (cluster centroid) $\mu_k$ in $\mathbb{R}^d$.

### K-means MacQueen [1967]

$$(\hat{\mu}_1, \ldots, \hat{\mu}_K, \hat{z}) \in \arg \min_{\mu_1, \ldots, \mu_K, z} \mathcal{J}(\mu_1, \ldots, \mu_K, z)$$

**objective function:**

$$\mathcal{J}(\mu_1, \ldots, \mu_K, z) = \sum_{k=1}^{K} \sum_{i=1}^{n} \| x_i - \mu_{z_i} \|^2$$
\textbf{K-means} \\

- a straightforward and widely used clustering algorithm, is one of the most important algorithms in unsupervised learning. \\
- Each cluster is represented by its mean (cluster centroid) \( \mu_k \) in \( \mathbb{R}^d \).

\textbf{K-means MacQueen [1967]}

\[
(\hat{\mu}_1, \ldots, \hat{\mu}_K, \hat{z}) \in \arg \min_{\mu_1, \ldots, \mu_K, z} J(\mu_1, \ldots, \mu_K, z)
\]

objective function: 
\[
J(\mu_1, \ldots, \mu_K, z) = \sum_{k=1}^{K} \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|^2
\]

- Initialization: \( (\mu_1^{(0)}, \ldots, \mu_K^{(0)}) \) (eg, randomly chosen data points)

1. **Assignment step**: 
   \[
   z_i^{(q)} = \arg \min_{z \in Z} \|x_i - \mu_z\|^2
   \]

2. **Relocation step**: 
   \[
   \mu_k^{(q+1)} = \frac{\sum_{i=1}^{n} z_{ik}^{(q)} x_i}{\sum_{i=1}^{n} z_{ik}^{(q)}}
   \]

\( \Rightarrow \) The \( K \)-means algorithm is simple to implement and relatively fast.
Example

Geyser Data

Faicel Chamroukhi  Statistical data science and some unsupervised learning problems  18/97
Geyser Data clustering K−means: K−means iteration: 1
Geyser Data clustering K-means: K-means iteration: 2

Faicel Chamroukhi
Statistical data science and some unsupervised learning problems
Example

Geyser Data clustering K-means: K-means iteration: 3

Faicel Chamroukhi
Statistical data science and some unsupervised learning problems
Example
Example

Geyser Data clustering K−means : K−means iteration : 5
Example
How to measure uncertainty?

**Figure** – $K$-means partition (left) vs GMM-EM partition (right).
Mixtures and the EM algorithm (Model-Based Clustering)

Finite Mixture Models [McLachlan and Peel., 2000]

\[ f(x; \theta) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k) \] with \( \pi_k > 0 \) \( \forall k \) and \( \sum_{k=1}^{K} \pi_k = 1 \).
Mixtures and the EM algorithm (Model-Based Clustering)

Finite Mixture Models [McLachlan and Peel., 2000]

\[ f(x; \theta) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k) \text{ with } \pi_k > 0 \forall k \text{ and } \sum_{k=1}^{K} \pi_k = 1. \]

Maximum-Likelihood Estimation

\[ \hat{\theta} \in \arg \max_{\theta} \ln L(\theta) \]

log-likelihood : \[ \ln L(\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(x_i; \theta_k). \]
Mixtures and the EM algorithm (Model-Based Clustering)

Finite Mixture Models [McLachlan and Peel., 2000]

\[ f(x; \theta) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k) \] with \( \pi_k > 0 \) \( \forall k \) and \( \sum_{k=1}^{K} \pi_k = 1 \).

Maximum-Likelihood Estimation

\[ \hat{\theta} \in \arg \max_{\theta} \ln L(\theta) \]

log-likelihood : \( \ln L(\theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(x_i; \theta_k) \).

The EM algorithm [Dempster et al., 1977]

\[ \theta^{new} \in \arg \max_{\theta \in \Omega} \mathbb{E} [\ln L_c(\theta)|\mathcal{D}, \theta^{old}] \]

complete log-likelihood : \( \ln L_c(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log [\pi_k f_k(x_i; \theta_k)] \)

where \( Z_{ik} \) is such that \( Z_{ik} = 1 \) if \( Z_i = k \) and \( Z_{ik} = 0 \) otherwise.
Mixtures and the EM algorithm (Model-Based Clustering)

Finite Mixture Models [McLachlan and Peel., 2000]

\[ f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}; \boldsymbol{\theta}_k) \text{ with } \pi_k > 0 \ \forall k \text{ and } \sum_{k=1}^{K} \pi_k = 1. \]

Maximum-Likelihood Estimation

\[ \hat{\boldsymbol{\theta}} \in \arg \max_{\boldsymbol{\theta}} \ln L(\boldsymbol{\theta}) \]

log-likelihood:
\[ \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k). \]

The EM algorithm [Dempster et al., 1977]

\[ \theta^{new} \in \arg \max_{\theta \in \Omega} \mathbb{E}[\ln L_c(\theta)|\mathcal{D}, \theta^{old}] \]

complete log-likelihood:
\[ \ln L_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log [\pi_k f_k(\mathbf{x}_i; \boldsymbol{\theta}_k)] \text{ where } Z_{ik} \]

is such that \( Z_{ik} = 1 \) if \( Z_i = k \) and \( Z_{ik} = 0 \) otherwise.

Clustering

\[ \hat{z}_i = \arg \max_{1 \leq k \leq K} \mathbb{P}(Z_i = k|\mathbf{x}_i; \hat{\boldsymbol{\theta}}), \quad (i = 1, \ldots, n) \]
Gaussian mixture models (GMMs)

The finite Gaussian mixture density is defined as:

\[ f(x_i; \Psi) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k) \]

**Figure** – An example of a three-component Gaussian mixture density in \( \mathbb{R}^2 \).
EM for Gaussian mixture models

1. **E-Step**: calculates the posterior component memberships:

\[
\tau_{ik}^{(q)} = \mathbb{P}(Z_i = k|x_i, \Psi^{(q)}) = \frac{\pi_k \mathcal{N}(x_i; \mu_k^{(q)}, \Sigma_k^{(q)})}{\sum_{\ell=1}^{K} \pi_{\ell} \mathcal{N}(x_i; \mu_{\ell}^{(q)}, \Sigma_{\ell}^{(q)})}
\]

that \(x_i\) originates from the \(k\)th component density.

2. **M-Step**: parameter updates:

\[
\pi_{k}^{(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)}}{n} = \frac{n_k^{(q)}}{n},
\]

\[
\mu_{k}^{(q+1)} = \frac{1}{n_{k}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} x_i,
\]

\[
\Sigma_{k}^{(q+1)} = \frac{1}{n_{k}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} (x_i - \mu_{k}^{(q+1)}) (x_i - \mu_{k}^{(q+1)})^T.
\]
Examples

Figure – A three-class example of a real data set: Iris data of Fisher.
Figure – Iris data: Clustering results with EM for a GMM and AIC.
Figure — Iris data of Fisher: The data are colored according to the true partition.
Iris Data clustering GMM : EM iteration : 0
Example
Iris Data clustering GMM : EM iteration : 2
Example

Iris Data clustering GMM: EM iteration: 3
Example

Iris Data clustering GMM : EM iteration : 4
Example
Example
Example
Iris Data clustering GMM : EM iteration : 9
Example

Iris Data clustering GMM : EM iteration : 10
Example

Iris Data clustering GMM : EM iteration : 11
Example
Example

Iris Data clustering GMM: EM iteration: 13
Example

Iris Data clustering GMM : EM iteration : 14
Example

Iris Data clustering GMM : EM iteration : 16

Faicel Chamroukhi
Statistical data science and some unsupervised learning problems
Example

Iris Data clustering GMM : EM iteration : 17
Iris Data clustering GMM : EM iteration : 20
Example

Iris Data clustering GMM : EM iteration : 21
Parsimonious GMMs for high-dimensional data

- Parsimonious Gaussian mixture models\(^1\) are statistical models that allow for capturing a specific cluster shapes (e.g., clusters having the same shape or different shapes, spherical or elliptical clusters, etc).

- Eigenvalue decomposition of the cluster covariance matrices:

\[
\Sigma_k = \lambda_k D_k A_k D_k^T
\]

where

- \(\lambda_k\) represents the volume of the \(k\)th cluster (the amount of space of the cluster).
- \(D_k\) is a matrix with columns corresponding to the eigenvectors of \(\Sigma_k\) that determines the orientation of the cluster.
- \(A_k\) is a diagonal matrix, whose diagonal entries are the normalized eigenvalues of \(\Sigma_k\) arranged in a decreasing order and its determinant is 1. This matrix is associated with the shape of the cluster.

---

1. Banfield and Raftery [1993], ?
Parsimonious GMMs for high-dimensional data

\[ \lambda I \]
\[ \lambda_k I \]
\[ \lambda A \]
\[ \lambda_k A \]
\[ \lambda A_k \]
\[ \lambda D A_k D_k^T \]
\[ \lambda_k D A_k D_k^T \]
\[ \lambda_k A_k \]
\[ \lambda D A D_k^T \]
\[ \lambda_k D A D_k^T \]
Model selection

- The problem of choosing the number of clusters can be seen as a model selection problem.
- The model selection task consists of choosing a suitable compromise between flexibility so that a reasonable fit to the available data is obtained, and over-fitting.
- A common way is to use a criterion (score function) that ensure the compromise.

\[
\text{score(model)} = \text{error(model)} + \text{penalty(model complexity)}
\]

which will be minimized.
- Here the complexity of a model $\mathcal{M}$ is related to the number of its (free) parameters $\nu$. 
Model selection

- Akaike Information Criterion (AIC):
  \[ \text{AIC}(\mathcal{M}_m) = \ln L(\hat{\theta}_m) - \nu_m \]

- Bayesian Information Criterion (BIC):
  \[ \text{BIC}(\mathcal{M}_m) = \ln L(\hat{\theta}_m) - \frac{\nu_m \log(n)}{2} \]

- Integrated Classification Likelihood (ICL):
  \[ \text{ICL}(\mathcal{M}_m) = \ln L_c(\hat{\theta}_m) - \frac{\nu_m \log(n)}{2} \]

where \( \ln L_c(\hat{\theta}_m) \) is the complete-data log-likelihood for the model \( \mathcal{M}_m \) and \( \nu_m \) denotes the number of free model parameters. For example, in the case of a \( d \)-dimensional Gaussian mixture model we have:

\[
\nu = (K - 1) + K \times d + K \times \frac{d \times (d + 1)}{2} = \frac{K \times (d + 1) \times (d + 2)}{2} - 1.
\]
Figure – Clustering results obtained with $K$-means algorithm (left) with $K = 2$ and the EM algorithm (right). The cluster centers are shown by the red and blue crosses and the ellipses are the contours of the Gaussian component densities at level $0.4$ estimated by EM. The number of clusters for EM have been chosen by BIC for $K = 1, \ldots, 4$. 
Examples

Figure — A three-class example of a real data set: Iris data of Fisher.
Figure – Iris data: Clustering results with EM for a GMM and AIC.
Figure – Iris data of Fisher: The data are colored according to the true partition.
Latent data models for dimensionality reduction

- Dimensionality reduction for high dimensional data (for representation/visualization etc)
- Principal Component Analysis (PCA) [Pearson, 1901, Hotelling, 1933],
- Factor Analysis (FA) [Spearman, 1904, Thurstone, 1947],
**Principal Component Analysis (PCA)**

- PCA is a linear projection which maximizes the variance in the projected space [Hotelling, 1933].

Consider a sample \( X = (x_1, \ldots, x_n) \) with \( x_i \in \mathbb{R}^d \).

⇒ The aim is to project the data onto a space having dimensionality \( M < d \) while maximizing the variance of the projected data.

Consider the sample mean vector and the sample covariance matrix:
\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T.
\]

⇒ The variance of the projected data is therefore given by the scalar:
\[
v(u) = \frac{1}{n} \sum_{i=1}^{n} (u^T x_i - u^T \bar{x})(u^T x_i - u^T \bar{x})^T = u^T Su. \tag{1}
\]

The **principal axes** (the direction vectors) are then given by:
\[
u = \arg \max_{u \in \mathbb{R}^d} u^T Su \tag{2}
\]

subject to \( u^T u = 1 \) and \( u_j^T u_k = 0 \) for \( j \neq k \).
Disadvantage:

The absence of a probability density model and associated likelihood measure.

Deriving PCA from the perspective of density estimation would offer a number of important advantages, including the following:

- The likelihood measure allows comparison with other density models.
- We can derive EM for PCA and hence deal with possible missing values.
- Possibility to perform Bayesian inference (e.g., for model selection).
- Possibility of computing the posterior class probabilities if PCA is used to model the class-conditional densities in classification.
- The value of the probability density function would give a measure of the novelty of a new data point.
- PCA model could be extended to a mixture framework.

\[ \Rightarrow \text{Use Probabilistic Principal Component Analysis (PPCA)} \]
Probabilistic Principal Component Analysis (PPCA)


\[ x_i = Wz_i + \mu + \epsilon_i \]  
Observed data = linear transf. of \( z \) + additive Gaussian noise

\[ z_i \sim \mathcal{N}(0, \sigma^2 I) \]  
latent variables of the principal component subspace

\[ \epsilon \sim \mathcal{N}(0, I) \]  
zero-mean Gaussian noise

\[ x_i | z_i \sim \mathcal{N}(Wz_i + \mu, \sigma^2 I) \]  
conditional density for the observed data

\[ x_i \sim \mathcal{N}(\mu, WW^T + \sigma^2 I) \]  
marginal density for the observed data

\[ (W, \mu, \sigma^2) \to \]  
EM for PPCA

NB : for $\mu$, we get its closed form solution : $\hat{\mu} = \bar{x}$

Only $W$ and $\sigma^2$ are computed in an iterative way by EM

1. **E-step** : By using the old parameters values, compute

$$
E[z_i] = (W^T W + \sigma^2 I)^{-1} W^T (x_i - \bar{x}) \tag{3}
$$

$$
E[z_i z_i^T] = \sigma^2 (W^T W + \sigma^2 I)^{-1} + E[z_i] E[z_i]^T \tag{4}
$$

2. **M-step**

$$
W_{new} = \left[ \sum_{i=1}^{n} (x_i - \bar{x}) E[z_i]^T \right] \left[ \sum_{i=1}^{n} E[z_i z_i^T] \right]^{-1} \tag{5}
$$

$$
\sigma^2_{new} = \frac{1}{nd} \sum_{i=1}^{n} \left\{ \|x_i - \bar{x}\|^2 - 2E[z_i]^T W_{new}^T (x_i - \bar{x}) + \text{trace}(E[z_i z_i^T] W_{new} W_{new}^T) \right\} \tag{6}
$$

NB. Here $E[.]$ is actually $E[. | X, \{W, \mu, \sigma^2\}_{old}]$
Factor Analysis (FA) [Spearman, 1904, Thurstone, 1947]
FA is closely related to PPCA
The only difference is

\[ x_i | z_i \sim \mathcal{N}(Wz_i + \mu, \Psi) \]

conditional density for the observed data

\[ \Psi \] is a \( d \times d \) diagonal matrix; rather than

\[ x_i | z_i \sim \mathcal{N}(Wz_i + \mu, \sigma^2 I) \]

conditional density for the observed data

(isotropic covariance matrix).
**Factor Analysis (FA) II**

**Generative model**

\[ x_i = Wz_i + \mu + \epsilon_i \]  
Observed data = linear transf. of \( z \) + additive Gaussian noise

\[ z_i \sim \mathcal{N}(0, \Psi) \]  
latent variables of the principal component subspace

\[ \epsilon \sim \mathcal{N}(0, I) \]  
zero-mean Gaussian noise

\[
\begin{align*}
x_i \mid z_i & \sim \mathcal{N}(Wz_i + \mu, \Psi) \text{ conditional density for the observed data} \\
x_i & \sim \mathcal{N}(\mu, WW^T + \Psi) \text{ marginal density for the observed data}
\end{align*}
\]
EM for Factor Analysis

1. **E-step**

\[
\mathbb{E}[z_i] = (I + WT \Psi^{-1}W)^{-1}WT(\Psi^{-1}x_i - \bar{x}) \quad (7)
\]

\[
\mathbb{E}[z_i; z_i^T] = (I + WT \Psi^{-1}W)^{-1} + \mathbb{E}[z_i] \mathbb{E}[z_i]^T \quad (8)
\]

2. **M-step**

\[
W_{\text{new}} = \left[ \sum_{i=1}^{n} (x_i - \bar{x})\mathbb{E}[z_i]^T \right] \left[ \sum_{i=1}^{n} \mathbb{E}[z_i; z_i^T] \right]^{-1} \quad (9)
\]

\[
\Psi_{\text{new}} = \text{diag} \left\{ S - W_{\text{new}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[z_i] (x_i - \bar{x})^T \right\} \quad (10)
\]

NB. Here \( \mathbb{E}[.] \) is actually \( \mathbb{E}[. | X, \{W, \mu, \Psi\}_{\text{old}}] \)
Time series segmentation

1. Statistics and Data Science
2. Clustering of multivariate data
3. Unsupervised learning for dimensionality reduction
4. Time series segmentation
5. Clustering of functional data
6. Unsupervised Bayesian (non-)parametric learning
7. Model-Based Co-Clustering of Multivariate Functional Data
Regression with hidden logistic process

Let \( y = (y_1, \ldots, y_n) \) be a time series of \( n \) univariate observations \( y_i \in \mathbb{R} \) observed at the time points \( t = (t_1, \ldots, t_n) \) governed by \( K \) regimes.

### The Regression model with Hidden Logistic Process (RHLP) [1]

\[
y_i = \beta_{z_i}^T x_i + \sigma z_i \epsilon_i ; \quad \epsilon_i \sim \mathcal{N}(0, 1), \quad (i = 1, \ldots, n)
\]

\[
Z_i \sim \mathcal{M}(1, \pi_1(t_i; w), \ldots, \pi_K(t_i; w))
\]

Polynomial segments \( \beta_{z_i}^T x_i \) with \( x_i = (1, t_i, \ldots, t_i^p)^T \) with logistic probabilities

\[
\pi_k(t_i; w) = \mathbb{P}(Z_i = k | t_i; w) = \frac{\exp (w_{k1} t_i + w_{k0})}{\sum_{\ell=1}^K \exp (w_{\ell1} t_i + w_{\ell0})}
\]

\[
f(y_i | t_i; \theta) = \sum_{k=1}^K \pi_k(t_i; w) \mathcal{N}(y_i; \beta_k^T x_i, \sigma_k^2)
\]

- Both the mixing proportions and the component parameters are time-varying
- Parameter estimation via the EM algorithm: EM-RHLP
Parameter estimation via a the EM algorithm : EM-RHLP

- Parameter estimation via a the EM algorithm (EM-RHLP)
  
  **M-Step :** includes a weighted logistic regression problem $\leftrightarrow$ IRLS
  
  (and weighted polynomial regressions)

- EM-RHLP algorithm complexity : $O(I_{EM} I_{IRLS} K^3 p^3 n)$ (more advantageous than dynamic programming).

Time series approximation and segmentation

1. Approximation : a curve prototype
   
   $\hat{y}_i = \mathbb{E}[y_i | t_i; \hat{\theta}] = \sum_{k=1}^{K} \pi_k(t_i; \hat{w}) \hat{\beta}_k^T x_i$

   $\leftrightarrow$ The RHLP can be used as nonlinear regression model

   $y_i = f(t_i; \theta) + \epsilon_i$

   by covering functions of the form

   $f(t_i; \theta) = \sum_{k=1}^{K} \pi_k(t_i; w) \beta_k^T x_i$  \[3\]

2. Curve segmentation : $\hat{z}_i = \arg \max_k \mathbb{E}[z_i | t_i; \hat{w}] = \arg \max_k \pi_k(t_i; \hat{w})$

   Model selection : Application of BIC, ICL ($\nu_{\theta} = K(p + 4) - 2.$)
Application to temporal data modeling and segmentation

Original Impedance spectrum
Approximated Impedance spectrum

\[ \pi_{ik}(w) \]

Time (Second)

Power (Watt)

Frequency [Hz]

\[ \text{Im}(Z) \text{ [mOhm]} \]

\[ \log_{10} \]
Joint segmentation of multivariate time series

Multiple hidden process regression

- Data: \((y_1, \ldots, y_n)\) a time series of \(n\) multidimensional observations \(y_i = (y^{(1)}_i, \ldots, y^{(d)}_i)^T \in \mathbb{R}^d\) observed at instants \(t = (t_1, \ldots, t_n)\).

- Model \(y_i = B^T_{z_i} x_i + e_i\); \(e_i \sim \mathcal{N}(0, \Sigma_{z_i}), (i = 1, \ldots, n)\)

\(z = (z_1, \ldots, z)\) A latent process generating the data

\(\leftrightarrow\) Multiple regression with hidden logistic process: Multiple RHLP [6]

\(\leftrightarrow\) Multiple Hidden Markov model regression (MHMMR) [7]

Application to human activity time series

![MRHLP segmentation of acceleration data issued from three body-worn sensors](Data from the LISSI Lab/University of Paris 12)
Outline

1. Statistics and Data Science
2. Clustering of multivariate data
3. Unsupervised learning for dimensionality reduction
4. Time series segmentation
5. Clustering of functional data
6. Unsupervised Bayesian (non-)parametric learning
7. Model-Based Co-Clustering of Multivariate Functional Data
Functional data are increasingly frequent

[James and Hastie, 2001, James and Sugar, 2003]
[Ramsay and Silverman, 2005]
[Chamroukhi et al., 2010]
[Bouveyron and Jacques, 2011]
[Samé et al., 2011]
[Jacques and Preda, 2014]
[Bouveyron et al., 2018]
[Chamroukhi and Nguyen, 2018]
High-dimensional FDA by clustering/segmentation

Non-stationary time series/functions

Railway curves

Satellite waveforms

Objectives

- Curve clustering/classification (functional data analysis framework)
- Deal with the problem of regime changes $\rightarrow$ Curve segmentation
Functional data clustering
Functional data clustering
# Functional data analysis context

## Data

- The individuals are entire functions (e.g., curves, surfaces)
- A set of $n$ univariate curves $((x_1, y_1), \ldots, (x_n, y_n))$
- $(x_i, y_i)$ consists of $m_i$ observations $y_i = (y_{i1}, \ldots, y_{im_i})$ observed at the independent covariates, (e.g., time $t$ in time series), $(x_{i1}, \ldots, x_{imi})$

## Objectives: exploratory or decisional

1. Unsupervised classification (clustering, segmentation) of functional data, particularly curves with regime changes: [4] [9], [C11] [16]
2. Discriminant analysis of functional data: [2], [5]

## Functional data clustering/classification tools

- A broad literature (Kmeans-type, Model-based, etc)
  
  $\Rightarrow$ Mixture-model based cluster and discriminant analyzes
The functional mixture model:

\[ f(y|x; \Psi) = \sum_{k=1}^{K} \alpha_k f_k(y|x; \Psi_k) \]

- \( f_k(y|x) \) are tailored to functional data: can be polynomial (B-)spline regression, regression using wavelet bases etc, or Gaussian process regression, functional PCA
- \( \rightarrow \) more tailored to approximate smooth functions
- \( \rightarrow \) do not account for segmentation

Here \( f_k(y|x) \) itself exhibits a clustering property via hidden variables (regimes):

1. Piecewise regression model (PWR)
2. Regression model with a hidden Markov process (HMMR)
3. Regression model with hidden logistic process (RHLP)
Piecewise regression mixture model (PWRM) [9]

- A probabilistic version of the $K$-means-like approach of [Hébrail et al., 2010]

$$ f(y_i | x_i; \Psi) = \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}(y_{ij}; \beta_{kr}^T x_{ij}, \sigma_{kr}^2) $$

$I_{kr} = (\xi_{kr}, \xi_{k,r+1}]$ are the element indexes of segment $r$ for component $k$

- Simultaneously accounts for curve clustering and segmentation

Parameter estimation

1. Maximum likelihood estimation : EM-PWRM

2. Maximum classification likelihood estimation : CEM-PWRM

- a generalization of the $K$-means-like algorithm of Hébrail et al. [2010] :

M-step : includes weighted piecewise regressions $\leadsto$ dynamic programming

Complexity in $\mathcal{O}(I_{EM} K R n m^2 p^3)$ : An issue for large $m$

Curve clustering : $\hat{z}_i = \arg \max_k \tau_{ik}(\hat{\Psi})$ with $\tau_{ik}(\hat{\Psi}) = \mathbb{P}(Z_i | x_i, y_i; \hat{\Psi})$
Application to switch operation curves

Data set: \( n = 146 \) real curves of \( m = 511 \) observations. Each curve is composed of \( R = 6 \) electromechanical phases (regimes)

<table>
<thead>
<tr>
<th>Method</th>
<th>Intra-cluster Inertia</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM-GMM</td>
<td>721.46</td>
</tr>
<tr>
<td>EM-PRM</td>
<td>738.31</td>
</tr>
<tr>
<td>EM-PSRM</td>
<td>734.33</td>
</tr>
<tr>
<td>( K )-means-like</td>
<td>704.64</td>
</tr>
<tr>
<td>CEM-PWRM</td>
<td>703.18</td>
</tr>
</tbody>
</table>

Table – Estimated intra-cluster inertia for the switch curves.
The Topex/Poseidon radar satellite data contains \( n = 472 \) waveforms of the measured echoes, sampled at \( m = 70 \) (number of echoes). We considered the same number of clusters (twenty) and a piecewise linear approximation of four segments per cluster as in Hébrail et al. [2010].

CEM-PWRM clustering of the satellite data
Mixture of hidden logistic process regressions \[4\]

The mixture of regressions with hidden logistic processes (MixRHLP):

\[
f(y_i | x_i; \Psi) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{m_i} \sum_{r=1}^{R_k} \pi_{kr}(x_j; w_k) \mathcal{N}(y_{ij}; \beta_{kr}^T x_j, \sigma_{kr}^2)\]

\[
\pi_{kr}(x_j; w_k) = \mathbb{P}(H_{ij} = r | Z_i = k, x_j; w_k) = \frac{\exp \left( w_{kr0} + w_{kr1} x_j \right)}{\sum_{r'=1}^{R_k} \exp \left( w_{kr'0} + w_{kr'1} x_j \right)},
\]

Two types of component memberships:

\(\rightarrow\) cluster memberships (global) \(Z_{ik} = 1 \iff Z_i = k\)

\(\leftarrow\) regime memberships for a given cluster (local) : \(H_{ijr} = 1 \iff H_{ij} = r\)

MixRHLP deals better with the quality of regime changes

Parameter estimation via the EM algorithm: EM-MixRHLP

EM-MixRHLP has complexity in \(O(I_{EM}I_{IRLS}KR^3mp^3)\) (\(K\)-means type for piecewise regression is in \(O(I_{KM}KRnm^2p^3)\) \(\leftarrow\) EM-MixRHLP is computationally attractive for large values of \(m\) and moderate values of \(R\).
EM-MixRHLP clustering of simulated data
Functional Linear Discriminant Analysis \[8\]
Functional Mixture Discriminant Analysis \[5\]

<table>
<thead>
<tr>
<th>Approach</th>
<th>Classification error rate (%)</th>
<th>Intra-class inertia</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLDA-PR</td>
<td>11.5</td>
<td>$10.7350 \times 10^9$</td>
</tr>
<tr>
<td>FLDA-SR</td>
<td>9.53</td>
<td>$9.4503 \times 10^9$</td>
</tr>
<tr>
<td>FLDA-RHLP</td>
<td>8.62</td>
<td>$8.7633 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-PRM</td>
<td>9.02</td>
<td>$7.9450 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-SRM</td>
<td>8.50</td>
<td>$5.8312 \times 10^9$</td>
</tr>
<tr>
<td>FMDA-MixRHLP</td>
<td>6.25</td>
<td>$3.2012 \times 10^9$</td>
</tr>
</tbody>
</table>
Phonemes data

Phonemes data set used in Ferraty and Vieu [2003]

1000 log-periodograms (200 per cluster)

Figure – Original phoneme data and curves of the five classes: “ao”, “aa”, “yi”, “dcl”, “sh”.

EM-like clustering results for Phonemes

Phonemes data set used in Ferraty and Vieu [2003] \(^4\)

1000 log-periodograms (200 per cluster)

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimated (K)</th>
<th>Misc. error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM-PRM</td>
<td>5</td>
<td>14.29 %</td>
</tr>
<tr>
<td>EM-SRM</td>
<td>5</td>
<td>14.09 %</td>
</tr>
<tr>
<td>EM-bSRM</td>
<td>5</td>
<td>14.2 %</td>
</tr>
</tbody>
</table>

EM-like clustering results for yeast cell cycle data

- Time course Gene expression data as in Yeung et al. [2001] \(^5\)
- 384 genes expression levels over 17 time points.

**Figure** – EM-like clustering results with the bSRM model.

Rand index : 0.7914 which indicates that the partition is quite well defined.

5. [http://faculty.washington.edu/kayee/model/]
Outline

1. Statistics and Data Science
2. Clustering of multivariate data
3. Unsupervised learning for dimensionality reduction
4. Time series segmentation
5. Clustering of functional data
6. Unsupervised Bayesian (non-)parametric learning
7. Model-Based Co-Clustering of Multivariate Functional Data
Bayesian spatial spline regression with mixed-effects

- Data : \( ((x_1, y_1), \ldots, (x_n, y_n)) \) a sample of \( n \) surfaces \( y_i = (y_{i1}, \ldots, y_{im_i})^T \) and their spatial coordinates \( x_i = ((x_{i11}, x_{i12}), \ldots, (x_{imi1}, x_{imi2}))^T \).

- Propose regression and regression mixtures, with three additional features:
  1. Include random effects
  2. Models for spatial functional data
  3. A full Bayesian inference

Bayesian spatial spline regression with mixed-effects [Esann 2016, 13]

\[
y_i = S_i(\beta + b_i) + e_i, \quad e_i \sim \mathcal{N}(0, \sigma^2 I_{m_i}), \quad (i = 1, \ldots, n)
\]

- \( \beta \) : fixed-effects regression coefficients
- \( b_i \) : random subject-specific regression coefficients \( b_i \perp e_i \sim \mathcal{N}(0, \xi^2 I_{m_i}) \)
- \( S_i \) is a spatial design matrix.
Bayesian mixture of spatial spline regressions

Data: A sample of \( n \) surfaces \((y_1, \ldots, y_n)\) and their spatial covariates \((S_1, \ldots, S_n)\) issued from \( K \) sub-populations

- Bayesian mixture of spatial spline regression models with mixed-effects (BMSSR):

\[
f(y_i | S_i; \Psi) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y_i; S_i(\beta_k + b_{ik}), \sigma_k^2 I_{m_i})
\]

\(\rightarrow\) Useful for density estimation and model-based clustering of heterogeneous surfaces

Hierarchical prior from for the BMSSR

\[
\begin{align*}
\pi & \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K) \\
\beta_k & \sim \mathcal{N}(\mu_0, \Sigma_0) \\
b_{ik} | \xi_k^2 & \sim \mathcal{N}(0_d, \xi_k^2 I_d) \\
\xi_k^2 & \sim \mathcal{IG}(a_0, b_0) \\
\sigma_k^2 & \sim \mathcal{IG}(g_0, h_0).
\end{align*}
\]
Bayesian inference of the BMSSR

- For the BMSSR, the parameter $\Psi$ is augmented by the unknown components labels $\mathbf{z} = (z_1, \ldots, z_n)$

Bayesian inference of the BMSSR using Gibbs sampling

- Sample from the analytic full conditional distributions:

  $Z_i|\ldots \sim \mathcal{M}(1; \tau_{i1}, \ldots, \tau_{iK})$ with $\tau_{ik}(1 \leq k \leq K) = \mathbb{P}(Z_i = k|\mathbf{y}_i, \mathbf{S}_i; \Psi)$

  $\pi|\ldots \sim \mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$

  $\beta_k|\ldots \sim \mathcal{N}(\nu_0, V_0)$

  $b_{ik}|\ldots \sim \mathcal{N}(\nu_1, V_1)$

  $\sigma^2_k|\ldots \sim \mathcal{IG}(g_1, h_1)$

  $\xi^2_k|\ldots \sim \mathcal{IG}(a_1, b_1)$

- relabel the obtained posterior parameter samples if label switching by the $K$-means-like algorithm of [Celeux, 1999, Celeux et al., 2000].
Handwritten digit clustering using the BMSSR

- BMSSR applied on a subset of the ZIPcode data set (issued from MNIST)
- Each individual $y_i$ contains $m_i = 256$ observations
  A subset of 1000 digits randomly chosen from the test set

![Cluster mean images obtained by the BMSSR model with 12 mixture components.](image)

**Figure** – Cluster mean images obtained by the BMSSR model with 12 mixture components.

The best solution is selected in terms of the Adjusted Rand Index (ARI) values, which promotes a partition with $K = 12$ clusters (ARI : 0.5238).
Outline

1. Statistics and Data Science
2. Clustering of multivariate data
3. Unsupervised learning for dimensionality reduction
4. Time series segmentation
5. Clustering of functional data
6. Unsupervised Bayesian (non-)parametric learning
7. Model-Based Co-Clustering of Multivariate Functional Data
Dirichlet Process Parsimonious Mixtures

- Mixture models for multivariate data in a fully Bayesian framework
- Dirichlet Process and Parsimonious Mixtures [C5,6,8], [11]

Dirichlet Processes (DP)

$\text{DP}(\alpha, G_0)$ [Ferguson, 1973] is a distribution over distributions:

$$\tilde{\theta}_i | G \sim G; \quad G | \alpha, G_0 \sim \text{DP}(\alpha, G_0), \ i = 1, 2, \ldots$$

Pólya urn representation [Blackwell and MacQueen, 1973]

$$\tilde{\theta}_i | \tilde{\theta}_1, \ldots, \tilde{\theta}_{i-1} \sim \frac{\alpha}{\alpha + i - 1} G_0 + \sum_{k=1}^{K_{i-1}} \frac{n_k}{\alpha + i - 1} \delta \theta_k$$

DP places its probability mass on an infinite mixture of Dirac deltas

$$G = \sum_{k=1}^{\infty} \pi_k \delta \theta_k \quad \theta_k | G_0 \sim G_0, \ k = 1, 2, ..., \text{ with } \sum_{k=1}^{\infty} \pi_k = 1$$

$\hookrightarrow$ The generated parameters $\tilde{\theta}_i$ for a DP process exhibit a clustering property
DPM : Generative model

\[ \begin{align*}
G|\alpha, G_0 & \sim \text{DP}(\alpha, G_0) \\
\tilde{\theta}_i|G & \sim G \\
\tilde{x}_i|\tilde{\theta}_i & \sim f(.|\tilde{\theta}_i)
\end{align*} \]

Chinese Restaurant Process mixtures (Pitman, 2002; Samuel and Blei, 2012)

- Latent variables \((z_1, \ldots, z_n)\)

- Predictive distribution:

\[
p(z_i = k|z_1, \ldots, z_{i-1}; \alpha) = \frac{\alpha}{\alpha + i - 1} \delta(z_i, K_{i-1} + 1) + \sum_{k=1}^{K_{i-1}} \frac{n_k}{\alpha + i - 1} \delta(z_i, k).
\]

- Generative model:

\[
\begin{align*}
\tilde{z}_i|\alpha & \sim \text{CRP}(z_{\setminus i}; \alpha) \\
\tilde{\theta}_{z_i}|G_0 & \sim G_0 \\
\tilde{x}_i|\tilde{\theta}_{z_i} & \sim f(.|\tilde{\theta}_{z_i})
\end{align*}
\]
Implemented parsimonious models

<table>
<thead>
<tr>
<th>Decomposition</th>
<th>Model-Type</th>
<th>Prior</th>
<th>Applied to</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda I$</td>
<td>Spherical</td>
<td>$IG$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>$\lambda_k I$</td>
<td>Spherical</td>
<td>$IG$</td>
<td>$\lambda_k$</td>
</tr>
<tr>
<td>$\lambda A$</td>
<td>Diagonal</td>
<td>$IG$</td>
<td>each diagonal element of $\lambda A$</td>
</tr>
<tr>
<td>$\lambda_k A$</td>
<td>Diagonal</td>
<td>$IG$</td>
<td>each diagonal element of $\lambda_k A$</td>
</tr>
<tr>
<td>$\lambda^{DAD^T}$</td>
<td>General</td>
<td>$IW$</td>
<td>$\Sigma = \lambda^{DAD^T}$</td>
</tr>
<tr>
<td>$\lambda_k^{DAD^T}$</td>
<td>General</td>
<td>$IG$ and $IW$</td>
<td>$\lambda_k$ and $\Sigma = DAD^T$</td>
</tr>
<tr>
<td>$\lambda^{DA_k D^T \ast}$</td>
<td>General</td>
<td>$IG$</td>
<td>each diagonal element of $\lambda A_k$</td>
</tr>
<tr>
<td>$\lambda_k^{DA_k D^T \ast}$</td>
<td>General</td>
<td>$IG$</td>
<td>each diagonal element of $\lambda_k A_k$</td>
</tr>
<tr>
<td>$\lambda^{D_k A D^T_k}$</td>
<td>General</td>
<td>$IG$</td>
<td>each diagonal element of $\lambda A$</td>
</tr>
<tr>
<td>$\lambda_k^{D_k A D^T_k}$</td>
<td>General</td>
<td>$IG$</td>
<td>each diagonal element of $\lambda_k A$</td>
</tr>
<tr>
<td>$\lambda^{D_k A_k D^T_k \ast}$</td>
<td>General</td>
<td>$IG$ and $IW$</td>
<td>$\lambda$ and $\Sigma_k = D_k A_k D^T_k$</td>
</tr>
<tr>
<td>$\lambda_k^{D_k A_k D^T_k \ast}$</td>
<td>General</td>
<td>$IW$</td>
<td>$\Sigma_k = \lambda_k^{D_k A_k D^T_k}$</td>
</tr>
</tbody>
</table>

Bayesian inference using Gibbs sampling

- Posterior distribution for the component labels:
  \[
p(z_i = k|z_{-i}, X, \Theta, \alpha) \propto p(x_i|z_i; \Theta)p(z_i|z_{-i}; \alpha) \text{ with } p(z_i|z_{-i}; \alpha) \text{ the CRP prior}
\]

- Posterior distribution for the component parameters:
  \[
p(\theta_k|z, X, \Theta_{-k}, \alpha; H) \propto \prod_{i|z_i = k} p(x_i|z_i = k; \theta_k)p(\theta_k; H) \text{ with } p(\theta_k; H) : \text{ Prior distribution over } \theta_k
\]

Bayesian model comparison by using Bayes Factors

\[
BF_{12} = \frac{p(X|M_1)p(M_1)}{p(X|M_2)p(M_2)} \approx \frac{p(X|M_1)}{p(X|M_2)} \text{ with the Laplace-Metropolis approximation}
\]

\[
p(X|M_m) = \int p(X|\theta_m, M_m)p(\theta_m|M_m)d\theta_m \approx (2\pi)^{\nu_m/2} |\hat{H}|^{1/2} p(X|\hat{\theta}_m, M_m)p(\hat{\theta}_m|M_m)
\]
Clustering of benchmarks

Diabetes data set, Geyser data set, Crabs data set

$2 \log BF : \lambda_k D_k A D_k^T \ vs \ \lambda D_k A D_k^T = 199.58$ (Decisive)

$2 \log BF : \lambda D A D^T \ vs \ \lambda_k D_k A D_k^T = 5$ (Substantial)

$\log 2BF : \lambda_k D_k A D_k^T \ vs \ \lambda_k D A D^T = 36.08$ (Decisive)
Humpback whale song decomposition

- Real fully unsupervised problem
- Data: 8.6 minutes of a Humpback whale song recording (with MFCC)

Objectives

- Discovering “call units”, which can be considered as a whale “alphabet”
- Find a partition of the whale song into clusters (segments), and automatically infer the unknown number of clusters from the data.
Unsupervised decomposition of whale song signals

- Sound demo of Unit 5 DPPM $\lambda I : (\text{sec. 0}) (\text{sec. 12})$
Unsupervised decomposition of whale song signals

Sound demo of Unit 8 DPPM $\lambda I$ : (sec. 8) (sec. 10)
Unsupervised decomposition of whale song signals

Sound demo of Unit 4 DPPM $\lambda_k A$ : (sec. 1) (sec. 7)
Unsupervised decomposition of whale song signals

- Sound demo of Unit 8 DPPM $\lambda_k A$ : (sec. 6) (sec. 12)
Outline

1. Statistics and Data Science
2. Clustering of multivariate data
3. Unsupervised learning for dimensionality reduction
4. Time series segmentation
5. Clustering of functional data
6. Unsupervised Bayesian (non-)parametric learning
7. Model-Based Co-Clustering of Multivariate Functional Data
   - Model-based co-clustering
   - Co-clustering of multivariate functional data
Outline

1. Statistics and Data Science
2. Clustering of multivariate data
3. Unsupervised learning for dimensionality reduction
4. Time series segmentation
5. Clustering of functional data
6. Unsupervised Bayesian (non-)parametric learning
7. Model-Based Co-Clustering of Multivariate Functional Data
   - Model-based co-clustering
   - Co-clustering of multivariate functional data
High-dimensional functional data clustering

- Multivariate functional data are increasingly present
- e.g.: Data continuously recorded for different subjects from multiple subject’s sensors

Measurements collected from different network elements (transceivers, cells, sites...):

**Figure** – An example with $d = 30$ and $n = 20$ daily observations [Ben Slimen et al., 2016].
High-dimensional functional data clustering

Questioning

Clustering of highly multivariate functional data with two guidelines:

- (1) Mathematical guideline: warranty for estimation and selection
- (2) User guideline: keep a user-friendly meaning of the process

Both are important because clustering is a highly risky task...

Proposed answering

(1) Model-based co-clustering with (2) temporal curve segmentation

Novelty corresponds to combining both (1) and (2)
Simultaneous clustering of lines/indiv. ($Z$) and columns/var. ($W$)

Can be used as a way to reduce dimensionality (var. $\rightarrow W$)

**Figure** – Binary data set with $n = 500$, $d = 300$, $K = M = 3$
Latent block model for co-clustering

The Latent Block Model [Govaert and Nadif, 2013]

\[
f(X; \Psi) = \sum_{(z,w) \in Z \times W} \mathbb{P}(Z, W; \pi, \rho) f(X|Z, W; \theta)
\]

Hypotheses

- The latent variables \( Z \) and \( W \) are independent: \( \mathbb{P}(Z, W) = \mathbb{P}(Z)\mathbb{P}(W) \) and iid:
  \[
  \mathbb{P}(Z) = \prod_i \mathbb{P}(z_i) \text{ with } z_i \sim \text{Multinomial}(\pi_1, \ldots, \pi_K) \text{ where } \pi_k = \mathbb{P}(z_k = k)
  \]
  \[
  \mathbb{P}(W) = \prod_j \mathbb{P}(w_j) \text{ with } w_j \sim \text{Multinomial}(\rho_1, \ldots, \rho_M) \text{ where } \rho_\ell = \mathbb{P}(w_j = \ell)
  \]
- Conditional independence: \( x_{ij}|(z_i, w_j) \perp x_{ij'}|(z_i, w_j') \)
Latent block model for co-clustering

The Latent Block Model [Govaert and Nadif, 2013]

\[ f(X; \Psi) = \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \mathbb{P}(Z, W; \pi, \rho) f(X | Z, W; \theta) \]

Hypotheses

- The latent variables \( Z \) and \( W \) are independent: \( \mathbb{P}(Z, W) = \mathbb{P}(Z)\mathbb{P}(W) \) and iid:
  \[ \mathbb{P}(Z) = \prod_i \mathbb{P}(z_i) \text{ with } z_i \sim \text{Multinomial}(\pi_1, \ldots, \pi_K) \text{ where } \pi_k = \mathbb{P}(z_k = k) \]
  \[ \mathbb{P}(W) = \prod_j \mathbb{P}(w_j) \text{ with } w_j \sim \text{Multinomial}(\rho_1, \ldots, \rho_M) \text{ where } \rho_\ell = \mathbb{P}(w_j = \ell) \]
- Conditional independence: \( x_{ij} | (z_i, w_j) \perp x_{i'j'} | (z_i', w_{j'}) \)

\( \leftrightarrow \) binary data: binary [Govaert and Nadif, 2003, 2008, Keribin et al., 2012],
\( \leftrightarrow \) categorical data: multinomial [Keribin et al., 2014]
\( \leftrightarrow \) contingency table: Poisson [Govaert and Nadif, 2003, 2006, 2008]
\( \leftrightarrow \) continuous data: Gaussian [Lomet, 2012, Govaert and Nadif, 2013]
\( \leftrightarrow \) functional data: functional PCA + Gaussian, see further [Ben Slimen et al., 2016]
Inference for the latent block model

Variational block EM (VBEM) for maximum likelihood estimation and fuzzy co-clustering [Govaert and Nadif, 2006, 2008].


Bayesian inference [Keribin et al., 2012, 2014]: Bayesian latent block mixtures for binary data and categorical data & a variational Bayesian inference and Gibbs sampling.

Number of blocks estimation: ICL criterion [Lomet, 2012, Keribin et al., 2014]
Functional data modeling: “classical” approach

[Ramsay and Silverman, 2005] and many others

- Step 1: \((x, y)\) decomposed into a finite basis of function (B-spline...): 
  \[ Y_i(t) \approx \sum_{r=1}^{d} c_{ir} \phi_r(x_i(t)) \] with \(c\) estimated by OLS.

- Step 2: functional principal components analysis (PCA) which is performed as a usual PCA of the basis expansion coefficients \(c\) using a metric defined by the inner products between the basis functions.

- Step 3: set a probability distribution on \(c\), typically Gaussian.

It defines a distribution on \(c\) instead of \(y\)...
Alternatively, use a segmentation via generative piecewise polynomial regression modeling of $f(y|x)$ [Chamroukhi et al.]

→ Regression with Hidden Logistic Process (RHLP)
→ See formula later

It gives a distribution on $y$ and also a meaningful segmentation of the curve
RHLP for modeling different types of functions
Multivariate functional data co-clustering

[Chamroukhi and Biernacki, 2017]

- Data: \( Y = (y_{ij}) \) a data sample matrix of \( n \) individuals defined on a set \( I \) and \( d \) continuous functional variables defined on a set \( J \).
- Each variable \( y_{ij} \) is an univariate curve \( y_{ij} = (y_{ij}(t_1), \ldots, y_{ij}(t_{T_{ij}})) \) of \( T_{ij} \) observations \( y(t) \in \mathbb{R} \) linked to covariates \( x_{ij} = (x_{ij}(t_1), \ldots, x_{ij}(t_{T_{ij}})) \) at the points \( (t_1, \ldots, t_{T_{ij}}) \), typically a sampling time.
Embedding RHLP in co-clustering

[Chamroukhi and Biernacki, 2017]

- Functional Latent Block Model for Co-clustering:

\[
 f(Y|X; \Psi) = \sum_{(z,w) \in Z \times W} P(Z; \pi)P(W; \rho)f(Y|X, Z, W; \theta) \\
 = \sum_{(z,w) \in Z \times W} \prod_i \pi_k^{z_i} \prod_j \rho_\ell^{w_j} \prod_{i,j,k,\ell} f(y_{ij}|x_{ij}; \theta_{k\ell})^{z_{ik} w_{j\ell}}.
\]

with parameter vector \(\Psi = (\pi^T, \rho^T, \theta^T)^T\), where \(\pi = (\pi_1, \ldots, \pi_K)^T\), \(\rho = (\rho_1, \ldots, \rho_M)^T\), and \(\theta = (\theta_{11}, \ldots, \theta_{k\ell}, \ldots, \theta_{KM})^T\).
Parameter estimation: EM not feasible

- Requires the calculation of the posterior joint distribution
  \[ P(z_{ik}w_{j\ell} = 1 | y_{ij}, x_{ij}) \]

- does not factorize due to the conditional dependence on the observed curves of the row and the column labels


← We adopt this variational approximation in our context

Variational block EM algorithm

Variational approximation

\[ P(z_{ik}w_{j\ell} = 1 | y_{ij}, x_{ij}) \approx P(z_{ik} = 1 | y_{ij}, x_{ij}) \times P(w_{j\ell} = 1 | y_{ij}, x_{ij}) \]
Variational block EM algorithm

\[ P(z_{ik}w_{j\ell} = 1|y_{ij}, x_{ij}) \approx P(z_{ik} = 1|y_{ij}, x_{ij}) \times P(w_{j\ell} = 1|y_{ij}, x_{ij}) \]
Variational block EM algorithm

\[
P(z_{ik}w_{j\ell} = 1 | y_{ij}, x_{ij}) \approx P(z_{ik} = 1 | y_{ij}, x_{ij}) \times P(w_{j\ell} = 1 | y_{ij}, x_{ij})
\]

**Initialization**: start from an initial solution at iteration \(q = 0\), and then alternate at the \((q + 1)\)th iteration between the following variational E- and M- steps until convergence:

**VE Step** Estimate the variational approximated posterior memberships:

1. \(\tilde{z}_{ik}^{(q+1)} \propto \pi_k^{(q)} \exp\left(\sum_{j,\ell,t,r} \tilde{w}_{j\ell}^{(q)} \tilde{h}_{tr}^{(q)} \log\left[ \alpha_{k\ell r}(t; \xi_{k\ell}^{(q)}) \mathcal{N}\left(y_{ij}(t); \beta_{k\ell r}^T x_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)\right]\right)\)

2. \(\tilde{w}_{j\ell}^{(q+1)} \propto \rho_{\ell}^{(q)} \exp\left(\sum_{i,k,t,r} \tilde{z}_{ik}^{(q)} \tilde{h}_{tr}^{(q)} \log\left[ \alpha_{k\ell r}(t; \xi_{k\ell}^{(q)}) \mathcal{N}\left(y_{ij}(t); \beta_{k\ell r}^T x_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)\right]\right)\)

3. \(\tilde{h}_{tr}^{(q+1)} \propto \alpha_{k\ell r}^{(q)}(t; \xi_{k\ell}^{(q)}) \mathcal{N}\left(y_{ij}(t); \beta_{k\ell r}^{(q)T} x_{ij}(t), \sigma_{k\ell r}^{(q)^2}\right)\)

where:

- \(\tilde{z}_{ik} = P(z_{ik} = 1 | y_{ij}, x_{ij})\),
- \(\tilde{w}_{j\ell} = P(w_{j\ell} = 1 | y_{ij}, x_{ij})\),
- \(\tilde{h}_{tr} = P(h_{tr} = 1 | z_i, w_j, y_{ij}(t), x_{ij}(t))\)
Variational block EM algorithm

**M Step** update the parameters estimates $\theta^{(q+1)}$ given the estimated posterior memberships at the current iteration $q + 1$:

1. $\pi_k^{(q+1)} = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$
2. $\rho_{\ell}^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$
Variational block EM algorithm

**M Step** update the parameters estimates $\theta^{(q+1)}$ given the estimated posterior memberships at the current iteration $q + 1$:

1. $\pi^{(q+1)}_k = \frac{\sum_i \tilde{z}_{ik}^{(q+1)}}{n}$

2. $\rho^{(q+1)} = \frac{\sum_j \tilde{w}_{jl}^{(q+1)}}{d}$

The update of each block parameters $\theta_{k\ell}$ consists in a weighted version of the RHLP updating rules:

3. $\xi_{k\ell}^{(new)} = \xi_{k\ell}^{(old)} - \left[ \frac{\partial^2 F(\xi_{k\ell})}{\partial \xi_{k\ell} \partial \xi_{k\ell}^T} \right]^{-1} \frac{\partial F(\xi_{k\ell})}{\partial \xi_{k\ell}} |_{\xi_{k\ell}^{(old)} = \xi_{k\ell}}$ which is the IRLS maximisation of $F(\xi_{k\ell}) = \sum_{i,j,t} \tilde{z}_{ik} \tilde{w}_{jl} h_{tr}^{(q)} \log \alpha_{k\ell r}(t; \xi_{k\ell})$ w.r.t $\xi_{k\ell}$. 

Faicel Chamroukhi  
Statistical data science and some unsupervised learning problems 88/97
**Variational block EM algorithm**

**M Step** update the parameters estimates $\theta^{(q+1)}$ given the estimated posterior memberships at the current iteration $q + 1$:

1. $\pi^{(q+1)}_k = \frac{\sum_i \tilde{z}^{(q+1)}_{ik}}{n}$

2. $\rho^{(q+1)}_\ell = \frac{\sum_j \tilde{w}^{(q+1)}_{j\ell}}{d}$

The update of each block parameters $\theta_{k\ell}$ consists in a weighted version of the RHLP updating rules:

3. $\xi^{(new)}_{k\ell} = \xi^{(old)}_{k\ell} - \left[ \frac{\partial^2 F(\xi_{k\ell})}{\partial \xi_{k\ell} \partial \xi_{k\ell}^T} \right]^{-1} \frac{\partial F(\xi_{k\ell})}{\partial \xi_{k\ell}} \bigg|_{\xi_{k\ell} = \xi^{(old)}_{k\ell}}$ which is the IRLS maximisation of $F(\xi_{k\ell}) = \sum_{i,j,t} \tilde{z}^{(q)}_{ik} \tilde{w}^{(q)}_{j\ell} \tilde{h}^{(q)}_{tr} \log \alpha_{k\ell r}(t; \xi_{k\ell})$ w.r.t $\xi_{k\ell}$.

The regression parameters updates consist in analytic WLS problems:

4. $\beta^{(q+1)}_{k\ell r} = \left[ \sum_{i,j} \tilde{z}^{(q)}_{ik} \tilde{w}^{(q)}_{j\ell} X_{ij}^T \Lambda_{ijkr}^{(q)} X_{ij} \right]^{-1} \sum_{i,j} \tilde{z}^{(q)}_{ik} \tilde{w}^{(q)}_{j\ell} X_{ij}^T \Lambda_{ijkr}^{(q)} y_{ij}$

5. $\sigma^{2(q+1)}_{k\ell r} = \frac{\sum_{i,j} \tilde{z}^{(q)}_{ik} \tilde{w}^{(q)}_{j\ell} \| \sqrt{\Lambda_{ijkr}^{(q)}} (y_{ij} - X_{ij} \beta^{(q+1)}_{k\ell r}) \|^2}{\sum_{i,j} \tilde{z}^{(q)}_{ik} \tilde{w}^{(q)}_{j\ell} \text{trace}(\Lambda_{ijkr}^{(q)})}$ where $X_{ij}$ is the design matrix for the $i$th curve, $\Lambda_{ijkr}^{(q)}$ is the diagonal matrix whose diagonal elements are the posterior segment memberships $\{ \tilde{h}^{(q)}_{i\ell tr}; t = 1, \ldots, T_{ij} \}$. 
It is also possible to use the Classification EM (CEM) approximation of EM [Celeux and Govaert, 1992].

### Parameter estimation by an SEM algorithm: SEM-FLBM

- The SEM algorithm [Celeux and Diebolt, 1985] allows to overcome some drawbacks of the variational-EM algorithm, including its sensitivity to starting values; SEM does not use an approximation.
- Eg. SEM for latent block models for categorical data [Keribin et al., 2012, 2014]
- The formulas of VEM-FLBM and SEM-FLBM are essentially the same, except that we incorporate a stochastic step consisting of sampling binary indicator variables $z_{ik}$, $w_{j\ell}$ and $h_{tr}$ according to $\tilde{z}_{ik}$, $\tilde{w}_{j\ell}$ and $\tilde{h}_{tr}$. 
Source codes are/will be made available on github

Matlab/R/Python

https://github.com/fchamroukhi
Data science, Big-Data, AI

The way of the future!

Eg. In France: Interdisciplinary Institutes of Artificial Intelligence (3IA):

Interdisciplinary Institutes of Artificial Intelligence (3IA): the four selected projects

The results of the 3IA (Interdisciplinary Institute of Artificial Intelligence) call for expressions of interest were made public on 6 November 2018 by Frédérique Vidal, French Minister of Higher Education, Research and Innovation, and Mounir Mahjoubi, French Secretary of State for Digital Affairs. The projects of the Grenoble (MIAI@Grenoble-Alpes), Nice (3IA Côte d’Azur), Paris (PRAIRIE) and Toulouse (ANITI) sites have been selected. Inria participates in three of the four successful projects.

At the heart of the national AI strategy

Each of the Inria research centres took part in formulating 3IA Institute projects within the framework of the considerable mobilisation of the regional academic and industry ecosystems: they are part of the dynamic stimulated by the national plan on artificial intelligence announced by the French president following the report by Cédric Villani.
Data science models/algorithms

New problems (big data, etc) but ... classical methods?

Our Core Algorithms Remain the Same

- Regression, decision trees, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent since the first Data Miner Survey in 2007.

Question: What algorithms/analytic methods do you TYPICALLY use? (Select all that apply)
References


Y. Ben Slimen, S. Allio, and J. Jacques. Model-Based Co-clustering for Functional Data. HAL preprint hal-01422756, December 2016. URL https://hal.inria.fr/hal-01422756.


Thank you for your attention!