Functional Latent Block Model

for functional data co-clustering

François-Xavier JOLLOIS
Université Paris Descartes, France

joint work with
C. Bouveyron (Univ. Nice), L. Bozzi (EDF) & J. Jacques (Univ. Lyon)

The data

- electricity consumption measured by Linky meters for EDF
- 27 millions of customers / 730 daily consumption over 2 years

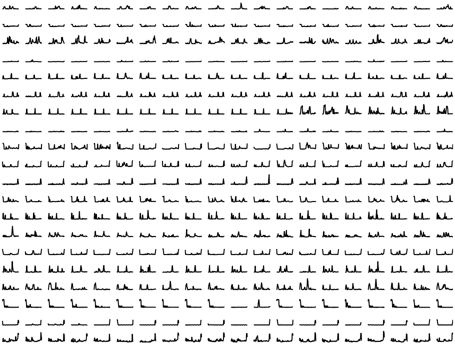


Figure: Sample of 20 consumptions for 20 days

The data

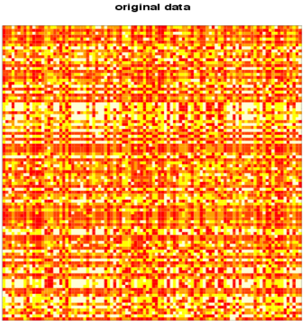
- large data matrix $\mathbf{x} = (x_{ij}(t))_{1 \leq i \leq n, 1 \leq j \leq p}$
- there is a need to summarize this data flow
- both n and p are (very) large

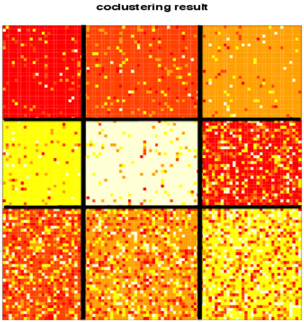
$\Rightarrow$ need for clustering of row (customers) and column (days of consumption):

need for co-clustering of functional data

Co-clustering ?

Simultaneous clustering of rows (individuals) and column (features)





legend: color level = $\frac{1}{T} \int_T x_{ij}(t)$

Electricity consumption = functional data

- $x_{ij}(t)$ are not totally known but only observed at a finite number of times points $x_{ij}(t_1), x_{ij}(t_2), \dots$
- need to reconstruct the functional nature of data

$\Rightarrow$ basis expansion assumption:

$$x_{ij}(t) = \sum_{h=1}^{m} a_{ijh}\phi_h(t), \quad t \in [0, T].$$

where $(\phi_h(t))_h$: spline, Fourier, wavelets...

- a_{ijh} estimated by least square smoothing

Overview

The fLBM model

Inference with SEM-Gibbs algorithm

Numerical experiments

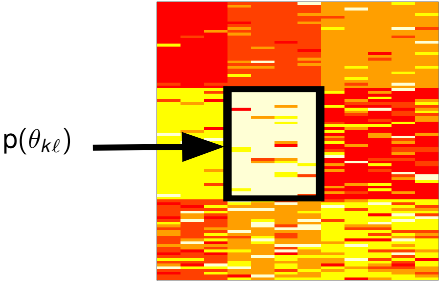
Application on EDF consumption curves

Plan

Latent Block Model (LBM)

Assumptions

- row $\mathbf{z} = (z_{ik})_{i,k}$ and column $\mathbf{w} = (w_{h\ell})_{h,\ell}$ partitions are independent
- conditionally on $(\mathbf{z}, \mathbf{w})$, x_{ij} are independent and generated by a block-specific distribution:



Latent Block Model

Latent Block Model (LBM)

$n \times d$ random variables $\mathbf{x}$ are assumed to be independent once the row $\mathbf{z} = (z_{ik})_{i,k}$ and column $\mathbf{w} = (w_{h\ell})_{h,\ell}$ partitions are fixed:

$$\mathrm{p}(\mathbf{x};\theta) = \sum_{\mathbf{z}\in V}\sum_{\mathbf{w}\in W} \mathrm{p}(\mathbf{z};\theta)\mathrm{p}(\mathbf{w};\theta)\mathrm{p}(\mathbf{x}|\mathbf{z},\mathbf{w};\theta)$$

with

- V (W) set of possible partitions of rows (column) into K (L) groups,
- $\mathrm{p}(\mathbf{z};\theta) = \prod_{ik} \alpha_k^{z_{ik}}$ and $\mathrm{p}(\mathbf{w};\theta) = \prod_{h\ell} \beta_\ell^{w_{h\ell}}$
- $\mathrm{p}(\mathbf{x}|\mathbf{z},\mathbf{w};\theta) = \prod_{ijk\ell} \mathrm{p}(\mathbf{a}_{ij};\theta_{k\ell})^{v_{ik}w_{h\ell}}$
- $\theta = (\alpha_k, \beta_\ell, \theta_{k\ell})$

The functional Latent Block Model (fLBM)

$p(\mathbf{a}_{ij};\theta_{k\ell})$ is the funHDDC distribution *(Bouveyron & Jacques, ADAC, 2011)*:

$$\mathbf{a}_{ij}|(z_{ik}=1,w_{j\ell}=1)\sim\mathcal{N}(U_{k\ell}\mu_{k\ell},U_{k\ell}\Sigma_{k\ell}U_{k\ell}^t+\Xi_{k\ell})$$

where

- $U_{k\ell}$ projects the $\mathbf{a}_{ij}$ into a low dimensional subspace for block $k\ell$
- $(\mu_{k\ell},\Sigma_{k\ell})$: (mean,variance) into the low-dimensional subspace,

$$Q_{k\ell}^t(U_{k\ell}\Sigma_{k\ell}U_{k\ell}^t+\Xi_{k\ell})Q_{k\ell}=\left(\begin{array}{cc}\begin{array}{|ccc|}\hline s_{k\ell 1} & & 0\\ & \ddots & \\ 0 & & s_{k\ell d}\\\hline\end{array} & \mathbf{0}\\ \mathbf{0} & \begin{array}{|ccc|}\hline b_{k\ell} & & 0\\ & \ddots & \\ 0 & & b_{k\ell}\\\hline\end{array}\end{array}\right)\begin{array}{l}\left.\vphantom{\begin{array}{c}1\\1\\1\end{array}}\right\}d\\\left.\vphantom{\begin{array}{c}1\\1\\1\end{array}}\right\}(m-d)\end{array}$$

with $s_{k\ell j}>b_{k\ell}$ for all $j=1,...,d$.

Plan

LBM inference

LBM inference

- The aim is to estimate θ by maximizing the observed log-likelihood

$$\ell(\theta; \mathbf{x}) = \sum_{\mathbf{v},\mathbf{w}} \ln \mathrm{p}(\mathbf{x}, \mathbf{v}, \mathbf{w}; \theta).$$

  where functional data $\mathbf{x}$ are represented by their coefficient $\mathbf{a}$, and $\mathbf{v}$ and $\mathbf{w}$ are missing row and column partitions
- EM is not computationally tractable
- $\Rightarrow$ variational or stochastic version should be used

SEM-Gibbs algorithm for LBM inference

- init : $\theta^{(0)}$, $\mathbf{w}^{(0)}$
- SE step
  - generate the row and column parititon $(\mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)})$ using a Gibbs sampling
- M step
  - Estimate θ, conditionally on $\mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)}$ obtained at the SE step.

SEM-Gibbs: SE step

1. generate the row partition $z_i^{(q+1)} = (z_{i1}^{(q+1)}, \ldots, z_{iK}^{(q+1)})|\mathbf{a}, \mathbf{w}^{(q)}$ for all $1 \leq i \leq n$ according to $z_i^{(q+1)} \sim \mathcal{M}(1, \tilde{z}_{i1}, \ldots, \tilde{z}_{iK})$ with for $1 \leq k \leq K$

$$\tilde{z}_{ik} = \mathrm{p}(z_{ik} = 1|\mathbf{a}, \mathbf{w}^{(q)}; \theta^{(q)}) = \frac{\alpha_k^{(q)} f_k(\mathbf{a}_i|\mathbf{w}^{(q)}; \theta^{(q)})}{\sum_{k'} \alpha_{k'}^{(q)} f_{k'}(\mathbf{a}_i|\mathbf{w}^{(q)}; \theta^{(q)})}$$

   where $\mathbf{a}_i = (\mathbf{a}_{ij})_j$ and $f_k(\mathbf{a}_i|\mathbf{w}^{(q)}; \theta^{(q)}) = \prod_{j\ell} \mathrm{p}(\mathbf{a}_{ij}; \theta_{k\ell}^{(q)})^{w_{j\ell}^{(q)}}$,

2. generate the column partition $w_j^{(q+1)} = (w_{j1}^{(q+1)}, \ldots, w_{jL}^{(q+1)})|\mathbf{a}, \mathbf{z}^{(q+1)}$ for all $1 \leq j \leq p$ according to $w_j^{(q+1)} \sim \mathcal{M}(1, \tilde{w}_{j1}, \ldots, \tilde{z}_{jL})$ with for $1 \leq \ell \leq L$

$$\tilde{w}_{j\ell} = \mathrm{p}(w_{j\ell} = 1|\mathbf{a}, \mathbf{z}^{(q+1)}; \theta^{(q)}) = \frac{\beta_\ell^{(q)} f_\ell(\mathbf{a}_j|\mathbf{z}^{(q+1)}; \theta^{(q)})}{\sum_{\ell'} \beta_{\ell'}^{(q)} f_{\ell'}(\mathbf{a}_j|\mathbf{z}^{(q+1)}; \theta^{(q)})}$$

   where $f_\ell(\mathbf{x}_j|\mathbf{z}^{(q+1)}; \theta^{(q)}) = \prod_{ik} \mathrm{p}(\mathbf{a}_{ij}; \theta_{k\ell}^{(q)})^{z_{ik}^{(q+1)}}$.

SEM-Gibbs: M step

same M step than for FunHDDC *(Bouveyron & Jacques, ADAC, 2011)*:

- $\alpha_k^{(q+1)} = \frac{1}{n}\sum_i z_{ik}^{(q+1)}$ and $\beta_\ell^{(q+1)} = \frac{1}{p}\sum_j w_{j\ell}^{(q+1)}$,
- $\mu_{k\ell}^{(q+1)} = \frac{1}{n_{k\ell}^{(q+1)}}\sum_i\sum_j \mathbf{a}_{ij}^{z_{ik}^{(q+1)} w_{j\ell}^{(q+1)}}$ with $n_{k\ell}^{(q+1)} = \sum_i\sum_j z_{ik}^{(q+1)} w_{j\ell}^{(q+1)}$,
- for the model parameters $s_{k\ell j}$, $b_{k\ell}$ and $Q_{k\ell j}$:
  - d first columns of Q_k: first eigenvectors of $\Omega^{\frac{1}{2}} C_{k\ell}^{(q)} \Omega^{\frac{1}{2}}$,
  - $s_{k\ell j}$, $j = 1, ..., d$: largest eigenvalues of $\Omega^{\frac{1}{2}} C_{k\ell}^{(q)} \Omega^{\frac{1}{2}}$,
  - b_k: $\mathrm{trace}(\Omega^{\frac{1}{2}} C_{k\ell}^{(q)} \Omega^{\frac{1}{2}}) - \sum_{j=1}^{d} s_{k\ell j}^{(q)}$,

  where $C_{k\ell}^{(q)}$ is the sample covariance matrix of block $k\ell$:

$$C_{k\ell}^{(q)} = \frac{1}{n_{k\ell}^{(q)}} \sum_{i=1}^{n}\sum_{j=1}^{p} z_{ik}^{(q+1)} \omega_{j\ell}^{(q+1)} (\mathbf{a}_{ij} - \mu_{k\ell}^{(q)})^t (\mathbf{a}_{ij} - \mu_{k\ell}^{(q)}),$$

and $\Omega = (\omega_{jk})_{1 \le j,k \le m}$ with $\omega_{jk} = \int_0^T \phi_j(t)\phi_k(t)dt$.

LBM inference

SEM-Gibbs algorithm for LBM inference

- $\hat{\theta}$ is obtained by mean of the sample distribution (after a burn in period)
- final bipartition $(\hat{\mathbf{v}}, \hat{\mathbf{w}})$ estimated by MAP conditionally on $\hat{\theta}$

LBM inference

Choosing K and L

We use the ICL-BIC criterion developed in (Lomet 2012) for continuous data co-clustering.
Thus, K and L can be chosen by maximizing

$$\text{ICL-BIC}(K, L) = \log \mathrm{p}(\mathbf{x}, \hat{\mathbf{v}}, \hat{\mathbf{w}}; \hat{\theta}) - \frac{K-1}{2} \log n - \frac{L-1}{2} \log p - \frac{KL\nu}{2} \log(np)$$

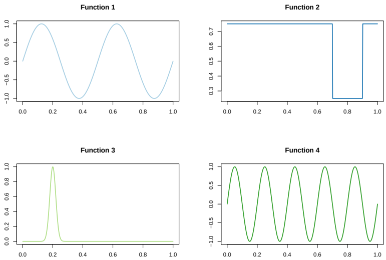
where $\nu = md + d + 1$ is the number of continuous parameters per block and

$$\log \mathrm{p}(\mathbf{x}, \hat{\mathbf{v}}, \hat{\mathbf{w}}; \hat{\theta}) = \prod_{ik} \hat{z}_{ik} \log \alpha_k + \prod_{j\ell} \hat{w}_{j\ell} \log \beta_\ell + \sum_{ijk\ell} \hat{z}_{ik} \hat{w}_{j\ell} \log \mathrm{p}(\mathbf{a}_{ij}; \hat{\theta}_{k\ell}).$$

Plan

Simulation setting

- $f_1(t), \ldots, f_4(t)$ are defined as block means



- all curves are sampled as follows:

$$x_{ij}(t)|Z_{ik}W_{jl} = 1 \sim \mathcal{N}(\mu_{kl}(t), \sigma^2),$$

  where $\sigma = 0.3$, $\mu_{11} = \mu_{21} = \mu_{33} = \mu_{42} = f_1$, $\mu_{12} = \mu_{22} = \mu_{31} = f_2$, $\mu_{13} = \mu_{32} = f_3$ and $\mu_{23} = \mu_{41} = \mu_{43} = f_4$.
- noise is added by adding $\tau\%$ of curves from other blocks.

3 scenarios of simulation

Table: Parameter values for the three simulation scenarios.

Scenario	A	B	C
n (nb. of rows)	100		
p (nb. of columns)	100		
T (length of curves)	30		
K (row groups nb.)	3	4	4
L (col. groups nb.)	3	3	3
α (row group prop.)	(0.333, ..., 0.333)	$(0.2, 0.4, 0.1, 0.3)$	$(0.2, 0.4, 0.1, 0.3)$
β (col. group prop.)	(0.333, ..., 0.333)	$(0.4, 0.3, 0.3)$	$(0.4, 0.3, 0.3)$
τ (simulation noise)	0	0.1	0.3

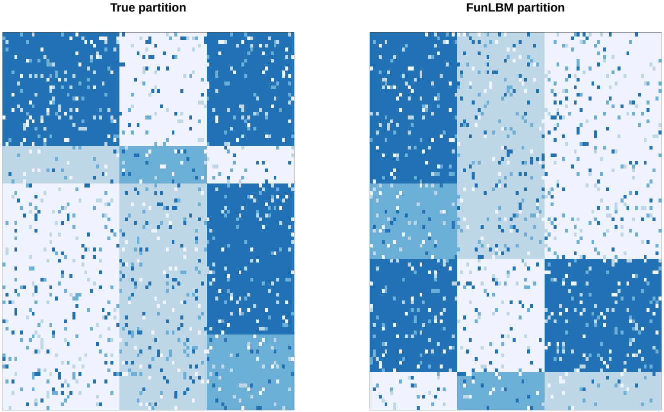
ICL performance for choosing (K, L)

Scenario A ($K = 3$, $L = 3$)						
$K \backslash L$	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	**100**	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

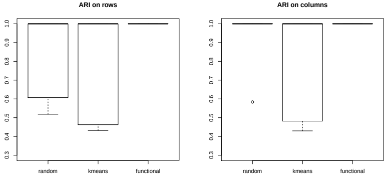
Scenario B ($K = 4$, $L = 3$)						
$K \backslash L$	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	**70**	0	1	0
5	0	0	26	1	0	0
6	0	0	2	0	0	0

Scenario C ($K = 4$, $L = 3$)						
$K \backslash Q$	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	0	17	0	0	0
3	0	0	**77**	0	0	0
4	0	0	5	0	0	0
5	0	0	1	0	0	0
6	0	0	0	0	0	0

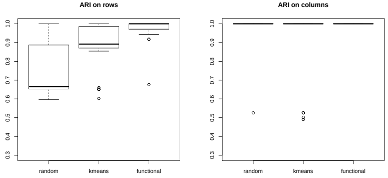
Co-clustering results for scenario B



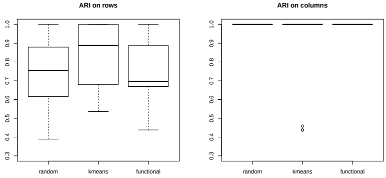
Co-clustering results



(a) scenario A



(b) scenario B



(c) scenario C

Figure: Adjusted Rand index values for the different initialization procedures on the three simulation scenarios.

Plan

The data

- electricity consumption measured by Linky meters for EDF
- 27 millions of customers / 730 daily consumption over 2 years

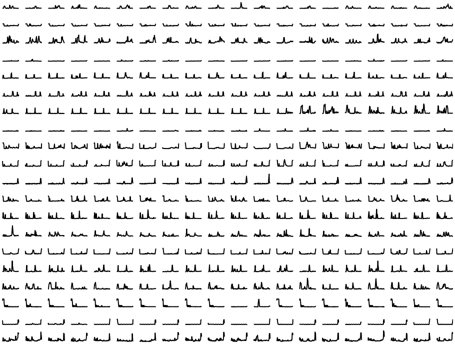
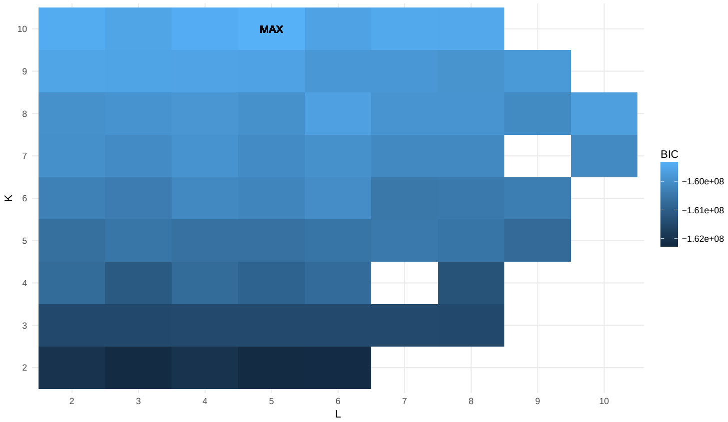
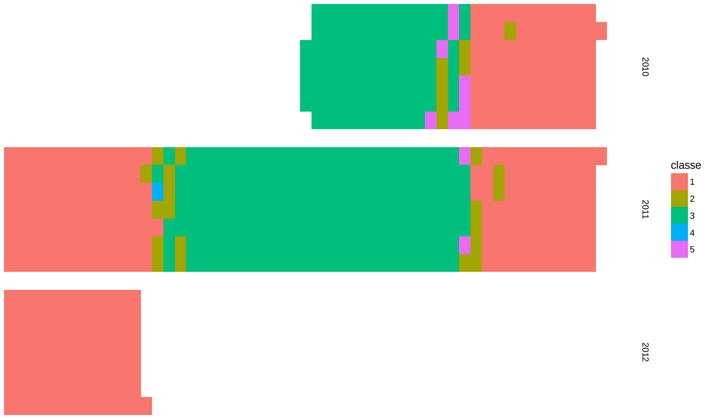


Figure: Sample of 20 consumptions for 20 days

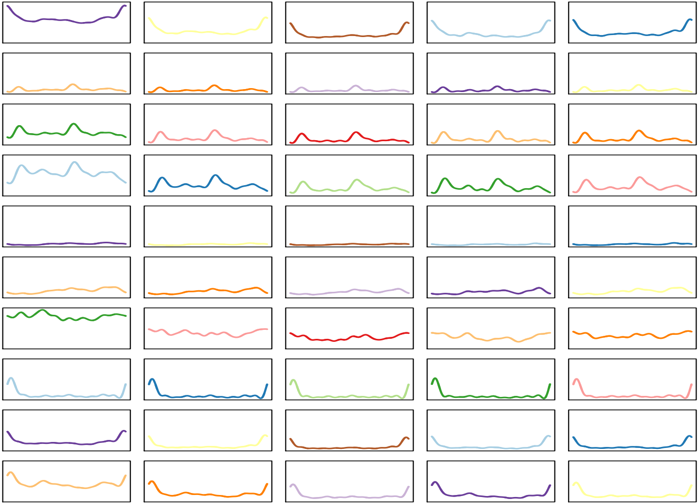
ICL values (choice of (K, L))



Clustering of columns (dates)



Average consumption curves of each block



Geographical clusters distributions

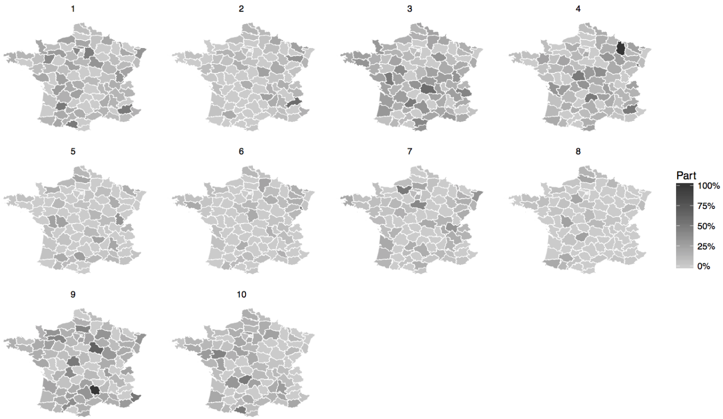


Figure: Proportions on households per French departments in each of the 10 clusters found by FunLBM.

Conclusions

Results

- real data application needs development of a co-clustering algorithm for functional data
- co-clustering algorithm has been developed based on a functional Latent Block model
- numerical experiments show the efficiency of SEM-Gibbs for model estimation as well as ICL-BIC for selecting of the number of blocks
- Results on EDF data are significant

References

- Bouveyron, C. and Jacques, J. (2011), Model-based Clustering of Time Series in Group-specific Functional Subspaces, Advances in Data Analysis and Classification, 5[4], 281-300.
- Govaert, G. and Nadif, M. (2013). Co-Clustering. Wiley-ISTE.