The Functional Latent Block Model for the Co-Clustering of Electricity Consumption Curves

Charles Bouveyron, Laurent Bozzi, Julien Jacques, François-Xavier Jollois

The Functional Latent Block Model for the Co-Clustering of Electricity Consumption Curves

Charles Bouveyron

Université Paris-Descartes, Laboratoire MAP5, UMR 8145, Paris, France.

Laurent Bozzi

EDF R&D, Paris-Saclay, France.

Julien Jacques

Université de Lyon, Lyon 2 & ERIC EA 3083, Lyon, France.

François-Xavier Jollois

Université Paris-Descartes, Laboratoire LIPADE, EA 2517, Paris, France.

Summary. As a consequence of the recent policies for smart meter development, electricity operators are nowadays able to collect data on electricity consumption widely and with a high frequency. This is in particular the case in France where EDF will be able soon to remotely record the consumption of its 27 millions clients every 30 minutes. We propose in this work a new co-clustering methodology, based on the functional latent block model (funLBM), which allows to build "summaries" of these large consumption data through co-clustering. The funLBM model extends the usual latent block model to the functional case by assuming that the curves of one block live in a low-dimensional functional subspace. Thus, funLBM is able to model and cluster large data set with high-frequency curves. An SEM-Gibbs algorithm is proposed for model inference. An ICL criterion is also derived to address the problem of choosing the number of row and column groups. Numerical experiments on simulated and original Linky data show the usefulness of the proposed methodology.

1. Introduction

Nowadays, electric meters are mostly electromechanical meters. They measure consumption and require a technician if a change in power or an outage occurs. Linky is a communicating meter, which means that it can receive and send data without the need for the physical presence of a technician. Installed in end-consumer properties and linked to a supervision centre, it is in constant interaction with the electricity network. After the installation of 300,000 smart meters "Linky" between 2009 and 2011 in the area of Lyon and Tours (France), the French authorities have decided to generalize these meters throughout the territory. By 2021, 35 million meters should be replaced in French households by Linky meters, allowing electricity operators to remotely record electricity consumption. For an operator like EDF with 27 millions of residential dwellings, these new smart meters represent a great opportunity to gather customer consumption data and therefore to improve client knowledge. Indeed, so far, customer data were recorded only every six months, while with the smart meter, the data can be taken up to every second. In practice, EDF plans to access the data every half hour, which means 17,472

measures per year for each of the 27 million customers. Nevertheless, this data flood may also be a drawback since they represent a mass of data to store and manage. To this end, it will be necessary to build meaningful "summaries" of these data, and one of the way to achieve that is though co-clustering.

The clustering of the time series corresponding to the customer consumptions, also quoted as *functional data* (Ramsay and Silverman, 2005), can be performed using functional data clustering techniques (see Jacques and Preda (2014) for a survey). With such approaches, the whole set of customers can be summarized into a small number of clusters. Nevertheless, the interpretation of these clusters, using for instance the mean consumption curves, is difficult due to the long period of observation (several months or even years). In order to provide a synthetic summary of the consumption data, it has been decided to cut the period of observation into small units of times. Taking into account EDF expectations in term of interpretation, the daily unit has been selected: thus, 365 daily curves consumption are observed per year for each of the 27 million customers. If for each cluster the interpretation of one mean daily consumption curve is feasible (and meaningful for EDF), it is still not possible to interpret them for all days. There is also a need to summarize the days of observations into a small number of clusters. Consequently, the analysis of the data provided by the Linky meters needs to build both clusters of customers and clusters of days of observation. From a statistical point of view, we are facing with a problem of clustering both the individuals (customers) and the features (days of observation), which is know in the literature as a *co-clustering* problem.

In the context of data recorded in a table where rows index individuals and columns index features, co-clustering techniques aims to simultaneously cluster individuals and features into homogeneous sets. Thus, the large data matrix can be summarized by a reduced number of blocks of data (or co-clusters). If the earliest (and most cited) method is probably due to Hartigan (1972), model-based approaches have recently proven their efficiency either for continuous, binary, categorical or contingency data (Govaert and Nadif, 2013; Jacques and Biernacki, 2017). Those latter approaches relies on the latent block model (LBM, Govaert and Nadif (2013)), which tackles with combinatorial issues by assuming local independence, *i.e.* all the random variables representing the cells of the data table are independent once the row and column partition have been fixed.

The originality of the present work is that the objects which have to be co-clusterize are functional data (electricity consumption curves). To the best of our knowledge, the only work dedicated to the co-clustering of such data is Ben Slimen et al. (2016), which proposes a co-clustering for functional data based on a two-steps approach: first, a functional PCA (fPCA, Ramsay and Silverman (2005)) is carried out onto the whole set of curves; second a Gaussian LBM is assumed onto the first fPCA scores. As model-based approaches have recently improved the two-steps approches in the clustering context (see Jacques and Preda (2014)), we propose in this work a functional Latent Block Model (funLBM) in order to improve the modeling capabilities of the model developed in Ben Slimen et al. (2016). The advantages of the funLBM model developed in this work are the following: first, the whole set of fPCA scores are modeled and not only the first ones as in Ben Slimen et al. (2016); second, the parametrization remains parsimonious since the data are assume to live in block-specific functional subspaces; finally, the fPCAs are carried out block per block, allowing to detect fine phenomena in the data structure.

The paper is organized as follows. Section 2 introduces the functional latent block model (funLBM) as well as its inference algorithm. Numerical experiments illustrates the interest and the behavior of the proposed co-clustering strategy in Section 3. Then, Section 4 presents the co-clustering analysis of the Linky meter data. Some concluding remarks are provided in Section 5.

2. The functional latent block model

After having introducing the notation, the funLBM model is defined. Then its inference is investigated through a Stochastic EM algorithm embedding a Gibbs sampling. The section ends with the definition of model selection criteria to choose the number of co-clusters.

2.1. The data

Let us consider that the data set is composed of a matrix of n customers (rows or samples) of p daily consumption curves (columns or functional features): $\boldsymbol{x} = (x_{ij}(t))_{1\leq i\leq n, 1\leq j\leq p}$ where $t \in [0, T]$ corresponds to the time in a day of observation ($T = 24$ for the Linky data). In practice, the functional expressions of the observed curves are not known and we only have access to the discrete observations at a finite set of ordered times. As explained in Aguilera et al. (2011), it is therefore necessary to first reconstruct the functional form of the data from their discrete observations. A common way to do this is to assume that curves belong to a finite dimensional space spanned by a basis of functions (see for example Ramsay and Silverman (2005)). Let also assume that each observed curve x_{ij} ($1 \leq i \leq n$, $1 \leq j \leq p$) can be expressed as a linear combination of basis functions $\{\phi_h\}_{h=1,\ldots,m}$:

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

The basis expansion coefficients $\boldsymbol{a}_{ij} = (a_{ijh})_h$ of each curve x_{ij} can be estimated by least square smoothing (Ramsay and Silverman, 2005). With this assumption, each curve x_{ij} will be represented by its basis expansion coefficient vector $\boldsymbol{a}_{ij}$. Let $\boldsymbol{a} = (\boldsymbol{a}_{ij})_{ij}$ be the whole data set to co-cluster.

2.2. The model

The latent block model (LBM, Govaert and Nadif (2013)) is certainly the most popular model for co-clustering. It assumes that the two random variables $\boldsymbol{z} = (z_{ik})_{1\leq i\leq n, 1\leq k\leq K}$ and $\boldsymbol{w} = (w_{j\ell})_{1\leq j\leq p, 1\leq \ell\leq L}$, indicating respectively the row and column partitions, are independent and that, conditionally to $\boldsymbol{z}$ and $\boldsymbol{w}$, the $n \times p$ random variables $\boldsymbol{x}$ are also independent. Note that standard binary partition is used for both $\boldsymbol{z}$ and $\boldsymbol{w}$, *i.e.* $z_{ik} = 1$ if observation i belongs to the row cluster k, 0 otherwise. Adapting it to functional data, we define the functional latent block model (funLBM):

$$\mathrm{p}(\boldsymbol{a};\theta) = \sum_{\boldsymbol{z}\in Z}\sum_{\boldsymbol{w}\in W} \mathrm{p}(\boldsymbol{z};\theta)\mathrm{p}(\boldsymbol{w};\theta)\mathrm{p}(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{w};\theta) \tag{1}$$

where (the straightforward ranges for i, j, k and ℓ are omitted hereafter):

- Z the set of all possible partitions of rows into K groups, W the set of partitions of the columns into L groups,
- $\mathrm{p}(\boldsymbol{z};\theta) = \prod_{ik} \alpha_k^{z_{ik}}$ and $\mathrm{p}(\boldsymbol{w};\theta) = \prod_{j\ell} \beta_\ell^{w_{j\ell}}$ where α_k and β_ℓ are the row and column mixing proportions, belonging to $[0,1]$ and summing to 1,
- $\mathrm{p}(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{w};\theta) = \prod_{ijk\ell} \mathrm{p}(\boldsymbol{a}_{ij};\theta_{k\ell})^{z_{ik}w_{j\ell}}$ is the probability density of the basis expansion coefficients $\boldsymbol{a}_{ij}$.

The model we assume for the density $\mathrm{p}(\boldsymbol{a}|\boldsymbol{z},\boldsymbol{w};\theta)$ is the one used for each cluster in the parsimonious FunHDDC model (Bouveyron and Jacques, 2011). We therefore assume that, for each block, there exists a low-dimensional latent subspace in which the curves can be adequately described. Following this model, $\mathrm{p}(\cdot;\theta_{k\ell})$ is a m-variate Gaussian density with mean $U_{k\ell}\mu_{k\ell}$ and variance $U_{k\ell}\Sigma_{k\ell}U_{k\ell}^t + \Xi_{k\ell}$:

$$\mathrm{p}(\boldsymbol{a}_{ij};\theta_{k\ell}) = \mathcal{N}(\boldsymbol{a}_{ij}; U_{k\ell}\mu_{k\ell}, U_{k\ell}\Sigma_{k\ell}U_{k\ell}^t + \Xi_{k\ell}),$$

where

- $U_{k\ell}$ is the $m \times d$ matrix ($d < m$) defined such that the orthogonal $m \times m$ matrix describing the linear transformation between the original space of the $\boldsymbol{a}_{ij}$ and the low-dimensional latent one can be decomposed into $Q_{k\ell} = [U_{k\ell}, V_{k\ell}]$ with $V_{k\ell}$ of size $m \times (m-d)$ with $U_{k\ell}^t U_{k\ell} = I_d$, $V_{k\ell}^t V_{k\ell} = I_{m-d}$ and $U_{k\ell}^t V_{k\ell} = 0$,
- $\mu_{k\ell}$ and $\Sigma_{k\ell}$ are the mean and variance of the projection of the basis expansion coefficients of the curves belonging to block $k\ell$ into the low-dimensional subspace, with $\Sigma_{k\ell} = diag(\sigma_{k\ell 1}, \ldots, \sigma_{k\ell d})$ a diagonal matrix,
- $\Xi_{k\ell}$ the noise covariance matrix of size $m \times m$, assuming to be such that $\Delta_{k\ell} = Q_{k\ell}^t(U_{k\ell}\Sigma_{k\ell}U_{k\ell}^t + \Xi_{k\ell})Q_{k\ell}$ can be written as follows:

$$\Delta_{k\ell} = \begin{pmatrix} \boxed{\begin{matrix} s_{k\ell 1} & & 0 \\ & \ddots & \\ 0 & & s_{k\ell d} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_{k\ell} & & 0 \\ & \ddots & \\ 0 & & b_{k\ell} \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} d \\ \left.\vphantom{\begin{matrix}a\\a\\a\end{matrix}}\right\} (m-d) \end{matrix}$$

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

- $\theta_{k\ell} = (\mu_{k\ell}, \Sigma_{k\ell}, U_{k\ell}, \sigma_{k\ell}^2)$ are the model parameter of block $k\ell$,

Let us finally denotes the whole set of mixture parameters by $\theta = (\alpha_k, \beta_\ell, \theta_{k\ell})_{1\le k\le K, 1\le \ell\le L}$.

2.3. Model inference with SEM-Gibbs algorithm

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

$$\ell(\theta; \boldsymbol{a}) = \sum_{\boldsymbol{z},\boldsymbol{w}} \ln \mathrm{p}(\boldsymbol{a}; \theta). \tag{2}$$

For computational reasons, and EM algorithm is not tractable in the co-clustering case (see Govaert and Nadif (2013)), thus we opt for one of its stochastic version, called SEM-Gibbs (Keribin C. and G., 2010). The main idea of this algorithm is, in a so-called SE step, to generate the unobserved row and column partitions $(\boldsymbol{z}, \boldsymbol{w})$ without having to compute their joint distribution (which is computationally intractable), thanks to a Gibbs sampling.

Starting from an initial value $\theta^{(0)}$ for the parameter set and an initial partition $\boldsymbol{w}^{(0)}$ for the unobserved column grouping, the qth iteration of the partial SEM-Gibbs alternates the following SE and M steps:

SE step. Execute a small number of iterations of the two following steps (Gibbs sampling):

(a) generate the row partition $z_i^{(q+1)} = (z_{i1}^{(q+1)}, \ldots, z_{iK}^{(q+1)})|\boldsymbol{a}, \boldsymbol{w}^{(q)}$ 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 $1 \leq k \leq K$

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

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

(b) generate the column partition $w_j^{(q+1)} = (w_{j1}^{(q+1)}, \ldots, w_{jL}^{(q+1)})|\boldsymbol{a}, \boldsymbol{z}^{(q+1)}$ 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 $1 \leq \ell \leq L$

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

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

M step. Estimate $\theta^{(q+1)}$ conditionally on $\boldsymbol{z}^{(q+1)}$ and $\boldsymbol{w}^{(q+1)}$. This can be done with the same M step than the one of the EM algorithm derived for FunHDDC inference (Bouveyron and Jacques, 2011). Mixture proportions are updated by $\alpha_k^{(q+1)} = \frac{1}{n}\sum_i z_{ik}^{(q+1)}$ and $\beta_\ell^{(q+1)} = \frac{1}{p}\sum_j w_{j\ell}^{(q+1)}$, whereas the block-means are updated as follows:

$$\mu_{k\ell}^{(q+1)} = \frac{1}{n_{k\ell}^{(q+1)}} \sum_i \sum_j \boldsymbol{a}_{ij}^{z_{ik}^{(q+1)} w_{j\ell}^{(q+1)}}$$

with $n_{k\ell}^{(q+1)} = \sum_i \sum_j z_{ik}^{(q+1)} w_{j\ell}^{(q+1)}$.

For the variance parameter updates, let us introduce the sample covariance matrix of the block indexed by $k\ell$:

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

and $\boldsymbol{\Omega} = (\Omega_{jk})_{1 \leq j,k \leq m}$, the matrix of inner products between the basis functions: $\Omega_{jk} = \int_0^T \phi_j(t)\phi_k(t)dt$. With these notations, the update formula for the model parameters $s_{k\ell j}$, $b_{k\ell}$ and $Q_{k\ell j}$ are:

- the d first columns of Q_k are updated by the eigenvectors associated with the largest eigenvalues of $\boldsymbol{\Omega}^{\frac{1}{2}} C_{k\ell}^{(q)} \boldsymbol{\Omega}^{\frac{1}{2}}$,
- the variance parameters $s_{k\ell j}$, $j = 1, ..., d$, are updated by the d largest eigenvalues of $\boldsymbol{\Omega}^{\frac{1}{2}} C_{k\ell}^{(q)} \boldsymbol{\Omega}^{\frac{1}{2}}$,
- the variance parameters b_k are updated by $\text{trace}(\boldsymbol{\Omega}^{\frac{1}{2}} C_{k\ell}^{(q)} \boldsymbol{\Omega}^{\frac{1}{2}}) - \sum_{j=1}^{d} s_{k\ell j}^{(q)}$.

The SEM-Gibbs algorithm is run for a given number of iterations (from our experiments, 3 iterations are sufficient to ensure a good behavior of the algorithm). After a burn-in period, the final estimation of the parameters is the mean of the sample distribution. Let us denote the final estimate by $\hat{\theta}$. Then, a sample of $(\boldsymbol{z}, \boldsymbol{w})$ is generated with the Gibbs sampling described above (SE step) with θ set to $\hat{\theta}$. The final bi-partition $(\hat{\boldsymbol{z}}, \hat{\boldsymbol{w}})$ is estimated by the mode of their sample distributions.

2.4. Model selection

Regarding model selection, since a model-based approach is proposed here, two funLBM models will be seen as different if they have different values of K and/or L . Therefore, the task of estimating K and L can be viewed as a model selection problem. Many model selection criteria have been proposed in the literature, such as the Akaike information criterion (AIC, Akaike (1974)) and the Bayesian information criterion (BIC, Schwarz (1978)). In this paper, because the optimization procedure considered involves binary matrices for $\boldsymbol{z}$ and $\boldsymbol{w}$, we rely on a ICL criterion. This criterion was originally proposed by (Biernacki et al., 2000) for Gaussian mixture models. We extend below to the functional case the ICL criterion developed by Lomet (2012) for co-clustering:

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

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

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

The couple (K, L) leading to the highest ICL value is selected as the most appropriate model for the data at hand.

Table 1. Parameter values for the three simulation scenarios (see text for details).

Scenario	A	B	C
n (nb. of rows)	100		
p (nb. of columns)	100		
T (length of curves)	30		
K (nb. of row groups)	3	4	4
L (nb. of column groups)	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 (column 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

3. Numerical experiments

This section aims at highlighting the main features of the proposed approach on synthetic data. In particular, the validity of the inference algorithm and model selection criterion, both presented in the previous section, is demonstrated on simulated data.

3.1. Simulation setup

To simplify the characterization and facilitate the reproducibility of the experiments, we designed a common simulation scenario on which the main characteristics of the proposed methodology will be illustrated. The simulation setup is as follows:

- we designed four different functions $f_1(t), ..., f_4(t)$, which will serve as block means, at equi-spaced time points $t = 0, 1/T, 2/T, ..., 1$. Figure 1 shows those functions.
- then, all curve points are sampled as follows:

$$x_{ij}(t)|\boldsymbol{z}_{ik}\boldsymbol{w}_{jl} = 1 \sim \mathcal{N}(m_{k\ell}(t), s^2),$$

  where $s = 0.3$, $m_{11} = m_{21} = m_{33} = m_{42} = f_1$, $m_{12} = m_{22} = m_{31} = f_2$, $m_{13} = m_{32} = f_3$ and $m_{23} = m_{41} = m_{43} = f_4$.
- finally, we add some noise within the blocks by randomly simulating a certain percentage τ of curves using other block means.

Table 1 provides the parameter values for the three simulation scenarios. Figure 2 shows a simulated data set according scenario B ($K = 4$, $L = 3$, noise $\tau = 0.1$) where colors indicate the used block mean functions. It is worth noticing that all simulation scenarios have been designed such that they do not follow the funLBM model and therefore they do not particularly favor our model in comparisons. In all the following experiments, Fourier basis functions are used to reconstruct the functional form of the data.

3.2. An introductory example

As an introductory example, we consider a data set simulated according to scenario B: $K = 4$ groups of rows, $L = 3$ group of columns, unbalanced row and column groups, and 10% of block noise. Figure 2 shows such a data set. In order to illustrate the behavior of the proposed inference algorithm, the SEM-Gibbs algorithm was run on the data with

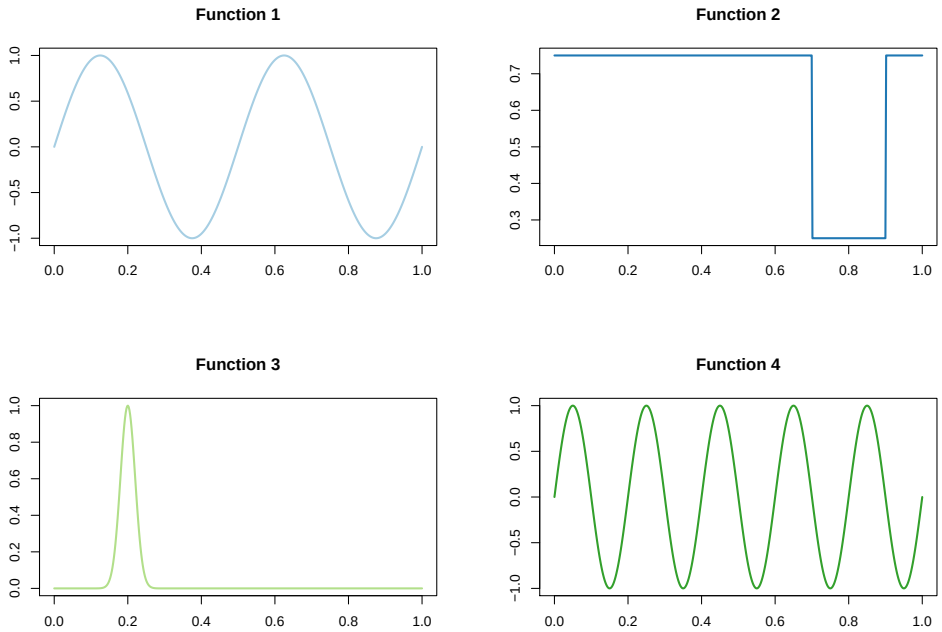


Fig. 1. The four block mean functions used in the simulations (see text for details).

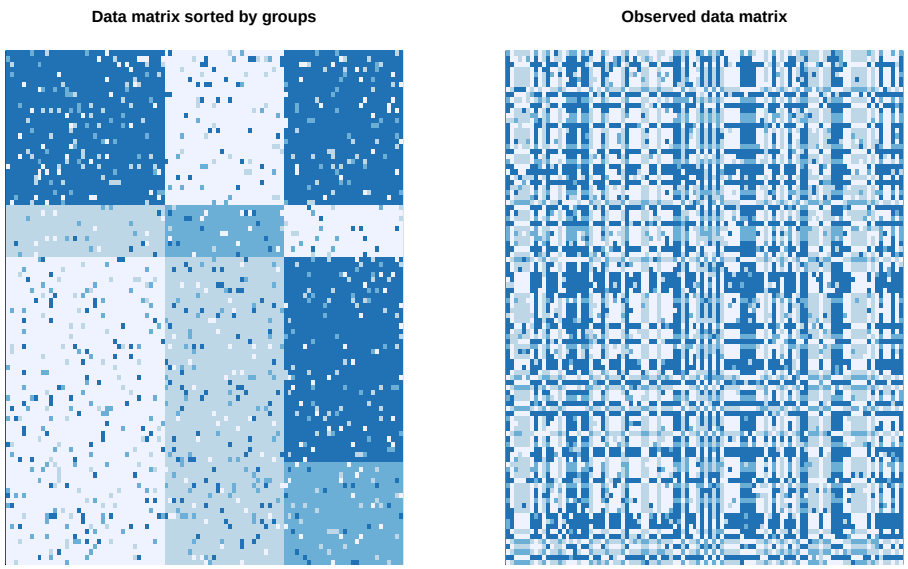
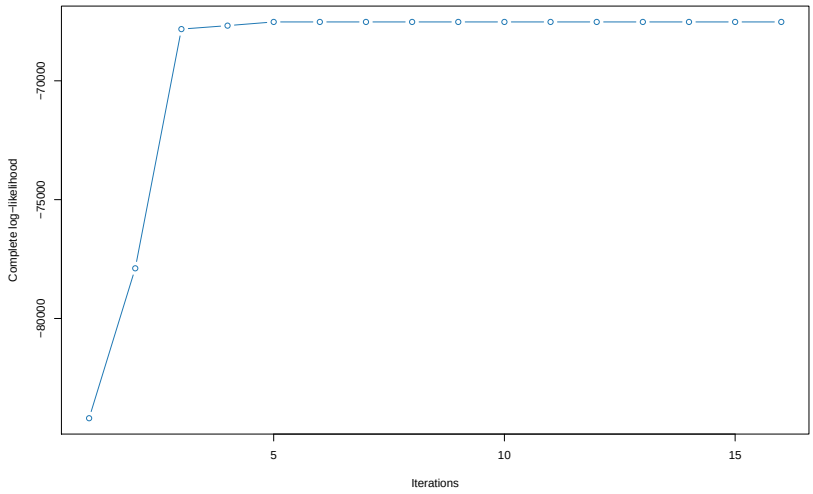
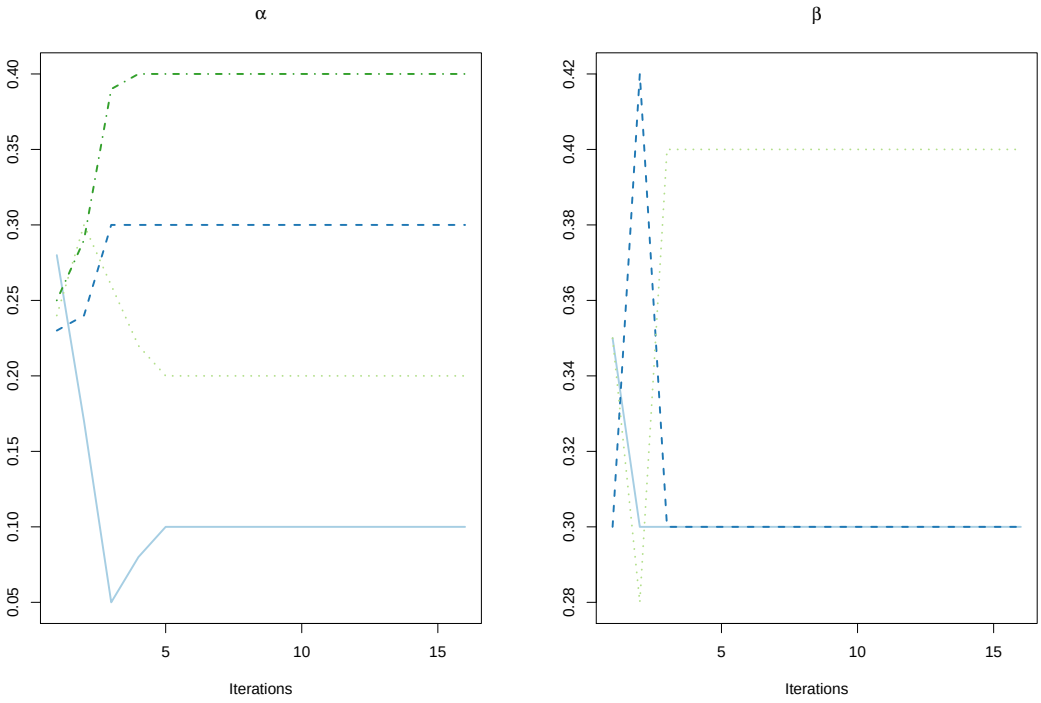


Fig. 2. A simulated data set with noise $\tau = 0.1$: colors indicate the used block mean functions.



(a) Complete data likelihood over the iterations of the funLBM algorithm.



(b) Estimates for mixture parameters

Fig. 3. Complete data likelihood (top) and estimates for mixture parameters (bottom) along the iterations of the SEM-Gibbs algorithm on the introductory example.

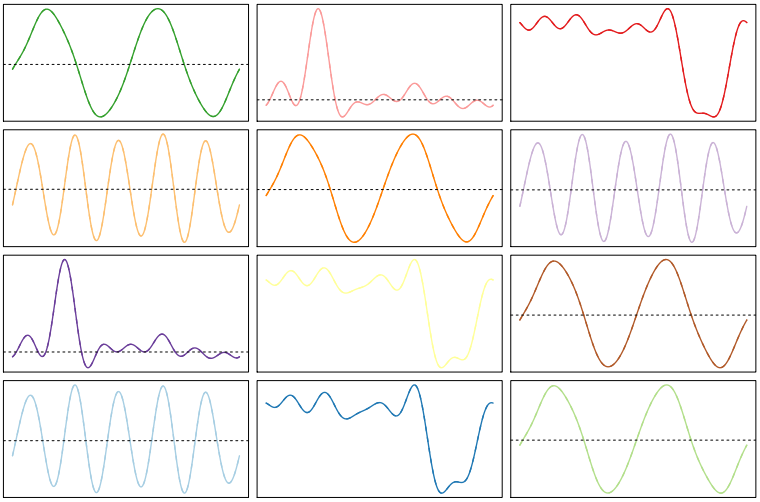


Fig. 4. Estimated functional means for the $K \times L$ blocks on the introductory example.

the actual numbers of row and column groups (the problem of model selection will be considered in next section). First, Figure 3 shows on the top panel the behavior of the complete data likelihood over the iterations of the funLBM algorithm. One can see that the funLBM algorithm converges in a few iterations. The figure also presents the evolution of the SEM-Gibbs estimates for model parameters α and β along the iterations. Figure 5 finally shows the obtained clustering, which is here perfect both regarding the simulated row and column partitions. Finally, it may be meaningful in practical cases to be able to visualize the functional means estimated by the funLBM algorithm. Figure 4 shows the estimated functional means for the $K \times L$ blocks and one may recognize functions very close to the block mean functions used in the simulations (Figure 1).

3.3. Initialization

We now focus on the initialization of SEM-Gibbs algorithm and consider three possible ways for that: random, kmeans and functional. The first possible initialization procedure is to randomly sample the $\boldsymbol{z}$ and $\boldsymbol{w}$ values. The second one consists in running the kmeans algorithm on the binning of the time series values according to the rows and then the columns. Finally, we also propose to use a functional clustering algorithm, funFEM (Bouveryon et al., 2015), instead of a kmeans on the binned functions.

We run our SEM-Gibbs algorithm with the three initialization strategies on 25 data sets simulated according to the three simulations scenarios. We evaluated the performance of the different results using the adjusted Rand index on both row and column partitions. In the clustering community, the adjusted Rand index (ARI, Rand (1971)) serves as a widely accepted criterion for the difficult task of clustering evaluation. The ARI looks at all pairs of nodes and check wether they are classified in the same group or not in both partitions. As a result, an ARI value close to 1 means that the partitions are similar and, in our case, that the funLBM algorithm succeeds in recovering the simulated

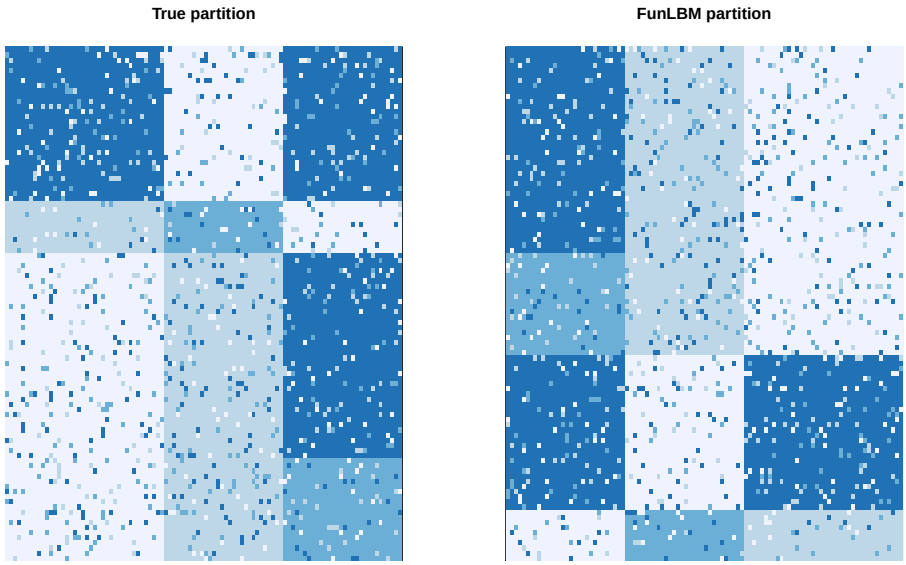


Fig. 5. Clustering results on the introductory example.

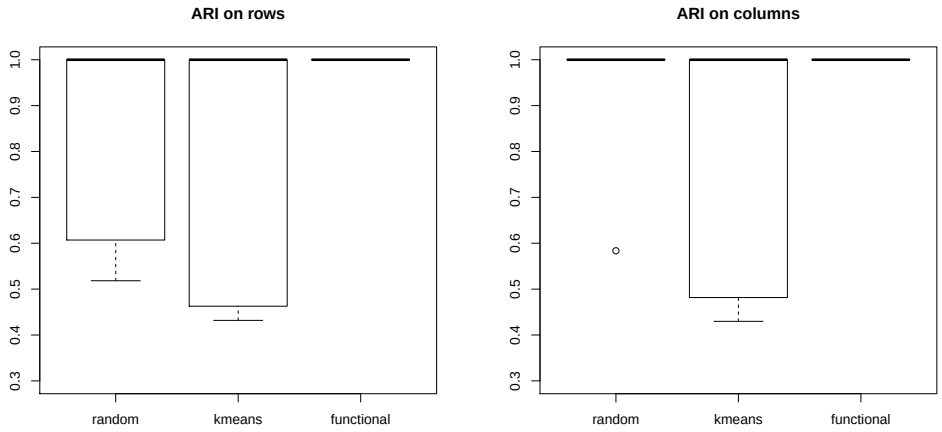
partitions. Figure 6 presents the results of this study.

As one can observe, the three techniques provide most of the time a satisfying result for both row and columns partitions on scenario A, which is the easier situation. For scenario B, which has unbalanced groups and some noise, the functional-based initialization clearly outperforms the two other strategy. Conversely, the kmeans initalization seems to be slightly superior on scenario C compared to the functional one. As a summary, the functional-based initialization may be viewed as the best overall solution and will be used in the following experiments.

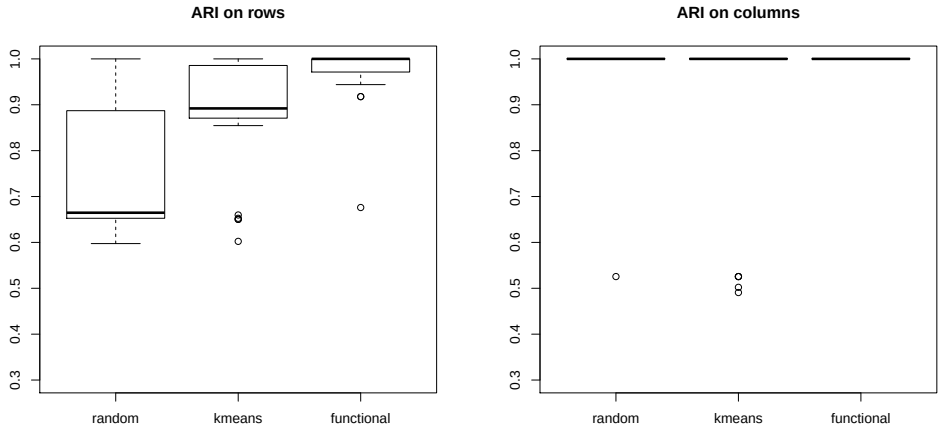
3.4. Model selection

This third simulation study focuses on model selection and aims at highlighting the ability of our approach to catch the actual model. To this end, 100 data sets were simulated for each scenario and the SEM-Gibbs algorithm (with the functional initialization) was applied in combination with our model selection criterion for values of K and L ranging from 1 to 6. Table 2 presents the percentage of selections by ICL for each model (K, L) on the 100 simulated data sets of each of the three scenarios.

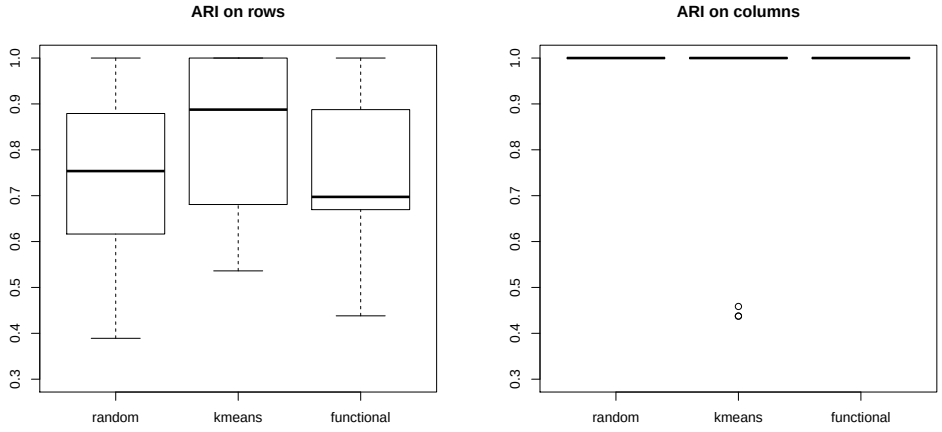
In the three different situations, our ICL criterion succeeds most of the time in identifying the actual combination of the number of row and column groups. For scenario A, the criterion allows our approach to identify perfectly the correct models. The task is of course slightly harder in the cases of the noisy scenarios B and C. The ICL allows nevertheless to recover the actual simulation model in more that 7 cases over 10. It is worth noticing that when ICL does not select the correct values for K and L, wrongly



(a) scenario A



(b) scenario B



(c) scenario C

Fig. 6. Adjusted Rand index values for the different initialization procedures on the three simulation scenarios.

Table 2. Percentage of selections by ICL for each model (K, L) on 100 simulated data sets of each of three scenarios. Highlighted rows and columns correspond to the actual values for K and L.

Scenario A ($K = 3$, $L = 3$)							Scenario B ($K = 4$, $L = 3$)							Scenario C ($K = 4$, $L = 3$)						
$K\backslash L$	1	2	3	4	5	6	$K\backslash L$	1	2	3	4	5	6	$K\backslash Q$	1	2	3	4	5	6
1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
2	0	0	0	0	0	0	2	0	0	0	0	0	0	2	0	0	17	0	0	0
3	0	0	**100**	0	0	0	3	0	0	0	0	0	0	3	0	0	**77**	0	0	0
4	0	0	0	0	0	0	4	0	0	**70**	0	1	0	4	0	0	5	0	0	0
5	0	0	0	0	0	0	5	0	0	26	1	0	0	5	0	0	1	0	0	0
6	0	0	0	0	0	0	6	0	0	2	0	0	0	6	0	0	0	0	0	0

selected models are usually close to the simulated one. Let us also recall that, since the data are not strictly simulated according to a funLBM model, the ICL criterion does not have the model which generated the data in the set of tested models. This experiment allows to validate ICL as a model selection tool for funLBM.

4. Application to EDF data

This section now presents the results of the modeling with funLBM of electricity consumption curves measured through autonomous meters, known as "Linky" meters.

4.1. Context of the study

With the upcoming installation of Linky meters in all french households, the field of operational applications of co-clustering methods is very wide at EDF. For instance, the co-clustering results may be used to design new marketing offers, to propose demand response programs or to detect outliers. Indeed, prior launching a new offer or service, some experimental trials are always made to evaluate the impact of the offer by comparing a test group and a control group. The two groups must be similar except the offer or the service tested. To ensure this hypothesis, we could select the samples among clusters built by the co-clustering technique. It is also possible to use the co-clustering results to design programs which consist in giving incentives to the customers to use less electricity at critical peak. Two ways exist: Price based demand response use changing prices to induce changes in customers consumption of electricity and direct load control, where equipment can be shut down remotely by the program operator. Once again, the groups of households and days found by the co-clustering may be used to parametrize these programs. Finally, with co-clustering techniques, the electricity operator should be able to detect outliers and warn customer if his consumption increases unusually, comparing to the mean or the quantiles of the clients of the same cluster.

4.2. Data and protocol

The Linky data set provided by EDF is made of 1,481 households in France (metropolitan) for which the electricity consumption has been monitored every 30 minutes and this over a period of almost two years (July 2010 – March 2012). The data set therefore consists in a table with $n = 1,481$ rows (households) and $p = 630$ columns (days),

Table 3. Selection of the most appropriate model for the Linky data.

Rank	K	L	ICL ($\times 10^6$)
1	**9**	**4**	**-116.09**
2	10	3	-116.11
3	9	3	-116.15
4	8	4	-116.18
5	8	3	-116.36

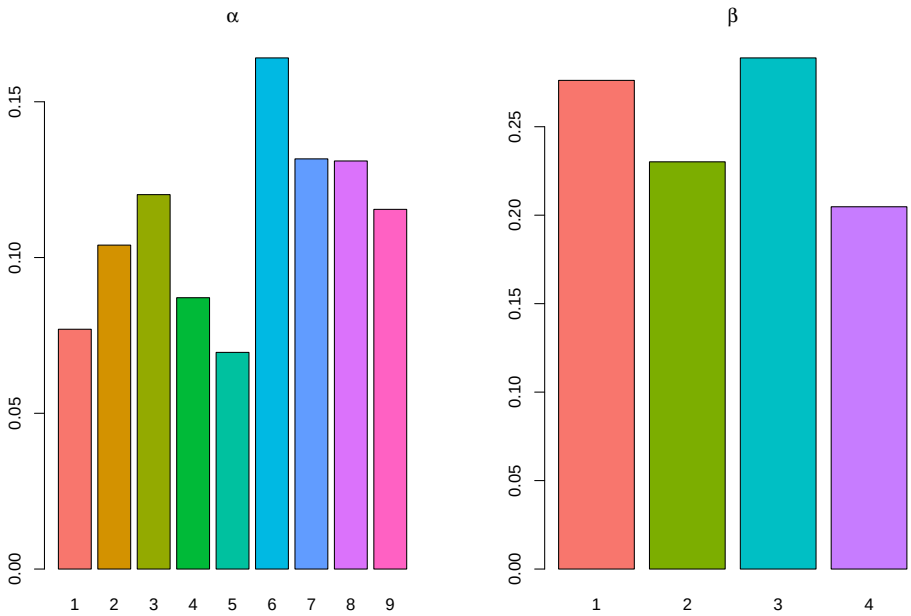


Fig. 7. Estimated proportions $\hat{\alpha}$ and $\hat{\beta}$ for the row (left) and column (right) groups.

where each entry is a time series of the daily electricity consumption with 48 measures. To transform the raw consumption data as meaningful functional data, the consumption data were first regressed against the observed temperatures at each household location to accommodate with geographic variations. Then, the residuals of such a regression were projected on a basis of 15 Fourier functions. We finally end up with a $1481 \times 630 \times 15$ cube which contains for each row and columns the 15 Fourier coefficients of the corresponding individual electricity consumption profiles. The SEM-Gibbs algorithm was run to infer funLBM models on those final data set for a number K of row (household) groups and a number L of column (date) groups ranging from 2 to 10.

4.3. Numerical results

As shown by Table 3, ICL selects the funLBM model with 9 groups of households and 4 groups of dates. Table 3 also shows the 5 models selected as the most appropriate ones for

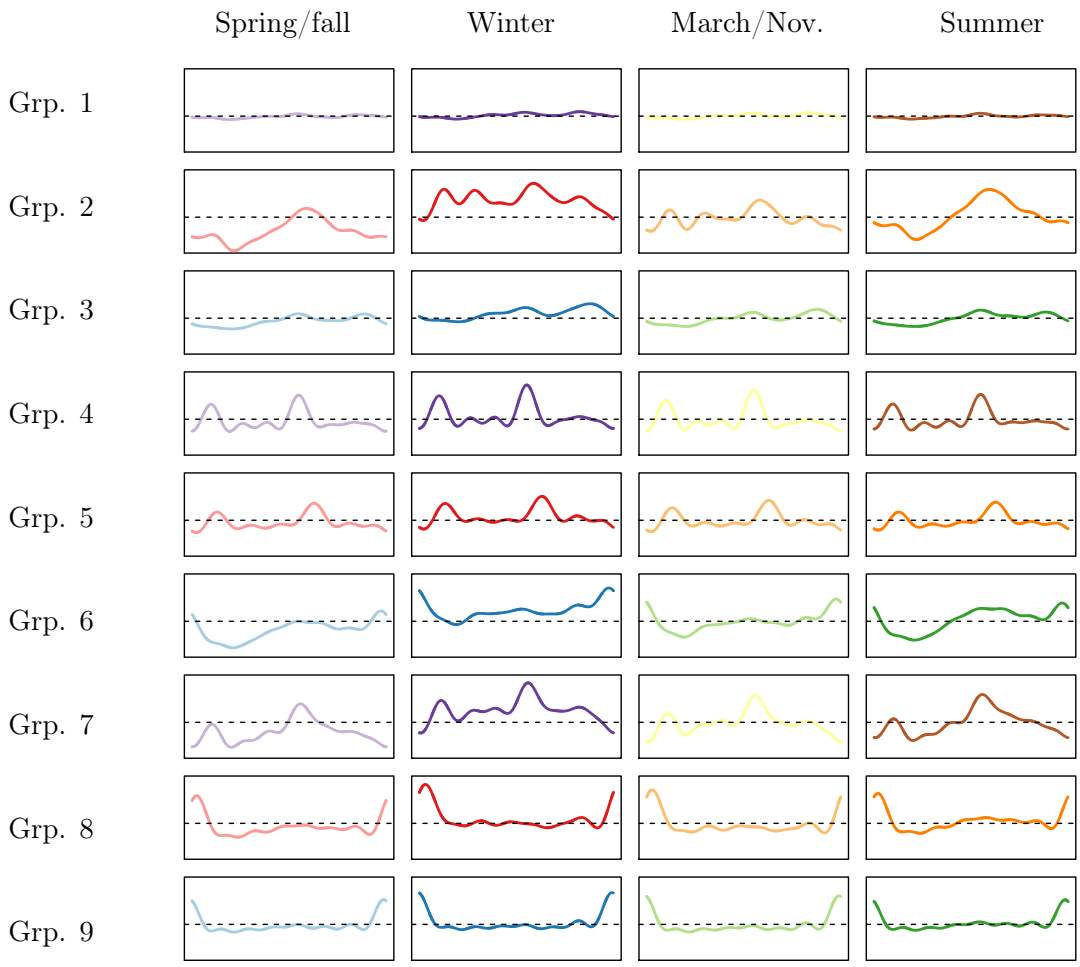


Fig. 8. Average curve of each block, as estimated by the SEM-Gibbs algorithm for funLBM.

the data by ICL. It is interesting to notice that, among the best models, there is a relative consensus on the choice of K and L, since those models are "centered" on the values of the best model. We therefore comment in the following the results corresponding to the funLBM model with $K = 9$ and $L = 4$.

Figure 7 shows the estimated proportions $\hat{\alpha}$ and $\hat{\beta}$ for respectively the row and column groups. It is interesting to notice that funLBM formed groups of rows and columns which are balanced, without extremely small or large groups. Figure 8 then presents the average consumption profile of the 9 groups of households for the 4 identified periods of time. Before to go further in the analysis, we have to explain that two major factors can have impacts on load curve profiles. First, the possession of electric heating system, which is particularly widespread in France (around 30 %), implies load peaks during winter and, in average, as soon as the external temperature falls down under 15 Celsius degrees. Second, the possession of an electric heating tank for sanitary water has also a significant

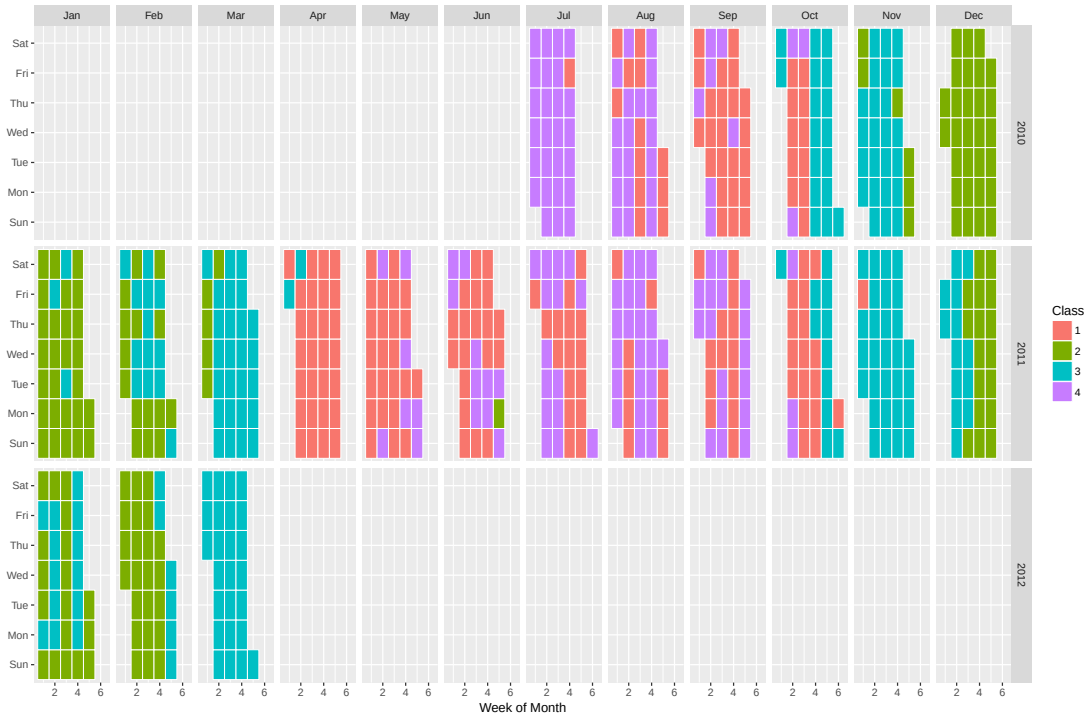


Fig. 9. Clustering of columns (dates) viewed as a calendar.

impact on load curve profiles. About 45% of the dwellings in France are equipped with such systems. To flatten its overall load curve, EDF was the first operator to create a time-of-use (TOU) tarification so that consumers benefit of 8 hours a day at a lower price to encourage them to shift some appliance use. Two main designs are proposed, one with 8 hours of off-peak prices during the night and another with 2 hours of off-peak price between noon and 5pm, and six remaining hours during the night.

At this point, it may be noticed from Figure 8 that some household groups have specific behaviors. On the one hand, the funLBM model allows to identify 6 groups (groups 1, 3, 4, 5, 8 and 9) of households which have stable electricity consumptions along the year, but which differ by their consumption profiles. These ones seem not have electricity as principal heating energy. For instance, the 1st and 3rd groups have an almost constant consumption of electricity both along the day and the year (they slightly differ in winter, probably by the use of a secondary heating system which may be electric), whereas the 4th and 5th groups have profiles which strongly vary within a day but the day profile is stable within a year. Their daily profile is typical from customers who benefit of off-peak hours during the day. The 8th and 9th groups reveals the daily behavior of customer with 8 hours of off-peak price in a row during the night.

On the other hand, 3 groups (groups 2, 6 and 7) have consumption profiles which are dependent of the time periods. To better understand those variations, it is necessary to have a look at the groups of days that funLBM provided. Figure 9 shows the group

memberships of the 630 days of the data set with a calendar view. In a few words, the 1st period (pink) corresponds to intermediate seasons (spring and fall), the 2nd period (green) gathers winter days, the 3rd one (blue) corresponds to the beginnings and endings of winter, and finally the 4th period (purple) is made of summer and spring holidays. In view of this interpretation of the four periods, the 6th and 7th household groups have similar consumption profiles but with different intensity: the consumption is higher in winter and summer than in intermediate seasons. More interestingly, the 2nd group (2nd top row of Figure 8) has significantly different profiles in winter and summer: in winter, they consume frequently during the day whereas, in summer, they consume only in the afternoon. It may due to the installation of a programmable controller for the office heating system.

Finally, let us highlight that one specific day has a surprising group membership: Wednesday June, 27th, 2011 is classified in the winter period. This may be explained by the fact that a heat wave was at its maximum on that day (37.5 Celsius degrees at Clermont-Ferrand), forcing people to use air conditioning at their maximum to cool their homes or offices. This unusual use of air conditioning may be viewed here as similar in term of electricity profiles to winter days where people use electric heating systems.

With this co-clustering method, EDF has obtained a very precise clustering of the load curves with clusters significantly different in terms of seasonality and of daily forms. For EDF, co-clustering with the funLBM model is an innovative approach since it allows to get these kind of clusters in one batch whereas EDF usually had to perform several independent steps. Indeed, the current approach at EDF for this is to combine two different clusterings based on dimension reduction (PCA and Kohonen), one for the seasonality, the other one for the daily forms. Let us also highlight that the funLBM framework allows to automatically identify the best combination of number K of household groups and number L of day groups.

5. Conclusion

To address the upcoming problem of exploring extremely large sets of electricity consumption curves, we proposed in this work a new co-clustering methodology for functional data, based on the functional latent block model (funLBM). The resulting co-clustering algorithm allows to build "summaries" of these large consumption data which can be efficiently used by operators. The funLBM model extends the LBM model to the functional case by assuming that the curves of one block live in a low-dimensional functional subspace. Model inference is done through a SEM-Gibbs algorithm and an ICL criterion has been also derived to address the model selection problem (choosing the number of row and column groups). Numerical experiments on simulated and original Linky data have shown the usefulness of the methodology.

Acknowledgments

This research has benefited from the support of the "FMJH Research Initiative Data Science for Industry" and from the support to this program from EDF and from Thalès Optronique.

References

Aguilera, A., M. Escabiasa, C. Preda, and G. Saporta (2011). Using basis expansions for estimating functional PLS regression. applications with chemometric data. *Chemometrics and Intelligent Laboratory Systems 104*(2), 289–305.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Ben Slimen, Y., S. Allio, and J. Jacques (2016). Model-based co-clustering for functional data. In *Proceedings of the 48th conference of the French Statistical Society*, Montpellier, France.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel 7*, 719–725.

Bouveryon, C., E. Côme, and J. Jacques (2015). The discriminative functional mixture model for the analysis of bike sharing systems. *Annals of Applied Statistics 9*(4), 1726–1760.

Bouveyron, C. and J. Jacques (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 M. Nadif (2013). *Co-Clustering.* Wiley-ISTE.

Hartigan, J. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association 67*(337), 123–129.

Jacques, J. and C. Biernacki (2017). Model-based co-clustering for ordinal data. Technical report, Preprint HAL 01448299.

Jacques, J. and C. Preda (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification 8*(3), 231–255.

Keribin C., G. G. and C. G. (2010). Estimation d'un modèle à blocs latents par l'algorithme sem. In *Proceedings of the 42th conference of the French Statistical Society*, Marseille, France.

Lomet, A. (2012). *Sélection de modèle pour la classification croisée de données continues.* Ph. D. thesis, Université de Technologie de Compiègne, Compiègne, France.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis* (Second ed.). Springer Series in Statistics. New York: Springer.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.