Model-Based Co-Clustering of Multivariate Functional Data

Faicel Chamroukhi
University of Caen, LMNO UMR CNRS 6139, France - faicel.chamroukhi@unicaen.fr

Christophe Biernacki
University of Lille, CNRS, INRIA - Lille, France - christophe.biernacki@inria.fr

Abstract

High dimensional data clustering is an increasingly interesting topic in the statistical analysis of heterogenous large-scale data. In this paper, we consider the problem of clustering heterogeneous high-dimensional data where the individuals are described by functional variables which exhibit a dynamical longitudinal structure. We address the issue in the framework of model-based co-clustering and propose the functional latent block model (FLBM). The introduced FLBM model allows to simultaneously cluster a sample of multivariate functions into a finite set of blocks, each block being an association of a cluster over individuals and a cluster over functional variables. Furthermore, the homogenous set within each block is modeled with a dedicated latent process functional regression model which allows its segmentation according to an underlying dynamical structure. The proposed model allows thus to fully exploit the structure of the data, compared to classical latent block clustering models for continuous non functional data, which ignores the functional structure of the observations. The FLBM can therefore serve for simultaneous co-clustering and segmentation of multivariate non-stationary functions. We propose a variational expectation-maximization (EM) algorithm (VEM-FLBM) to monotonically maximize a variational approximation of the observed-data log-likelihood for the unsupervised inference of the FLBM model.
Keywords: Co-clustering; Mixture modeling; Latent block model; Functional data analysis; Curve clustering; EM algorithms; Variational EM

1 Introduction and related work

High dimensional data clustering is an increasingly interesting topic in the statistical analysis of heterogenous large-scale data. Many statistical studies involve observations issued from underlying entire functions (i.e. curves). The most frequent case of functional representations is that in which the studied individuals have a temporal variability (i.e. time series). This "functional" aspect of the data adds additional difficulties in the analysis compared to the case of a classical multivariate (non functional) analysis, which ignores the underlying structure of the individuals. The adapted paradigm of analyzing such data is the increasing framework of functional data analysis (FDA) [23]. The key tenet of FDA is to treat the data not just as multivariate observations but as (discretized) values of underlying smooth functions (e.g. curves).
Here we consider the problem of the unsupervised analysis of heterogeneous high-dimensional functional data via clustering. One of the most popular and successful approaches in cluster analysis is model-based clustering (e.g. see [7]), that is the one based on the flexible and statistically sound framework of mixture models [21], and the well-known desirable properties of the expectation-maximization (EM) algorithm [5, 20]. This flexible modeling framework in multivariate analysis is taking a growing investigation for FDA. See for example [2, 6, 8, 13, 14, 15, 18].

In high-dimensional scenarios, one extended framework of model-based clustering framework to the clustering of individuals described by a large set of variables, is model-based co-clustering. While model-based clustering techniques aim at providing a partition of the data into homogeneous groups of individuals, or possibly in variables, model-based co-clustering [9, 11, 12], also called bi-clustering or block clustering, aim at simultaneously co-clustering the data into homogeneous blocks, a block being a simultaneous association of individuals and variables. They rely on latent block models [12] and have been developed for binary data [9, 11, 16], categorical data [17], contingency table [9, 10, 11] and continuous data [12, 19]. The block-mixture can be estimated by a block classification EM (CEM) algorithm for maximum classification likelihood and hard co-clustering [9, 10, 11], or a variational block EM (VBEM) for maximum likelihood estimation and fuzzy co-clustering [10, 11]. The block mixture models have also been examined from a Bayesian prospective to deal with some problems encountered in the MLE approach. Recently, [16, 17] proposed in the Bayesian formulation of the latent block mixture, for respectively binary data and categorical data, a variational Bayesian inference and Gibbs sampling technique.

However, these statistical analyses in model-based co-clustering are designed for multivariate vectorial data. For multivariate functional data issued from underlying continuous functions (e.g. curves), a standard classical multivariate co-cluster analysis is not well-adapted as it does not fully exploit the underlying structure of the individuals. We therefore consider the model-based co-clustering of functional data, which is much less explored compared to the previously described model-based clustering approaches. Indeed, it is only very recently that we start having models dedicated to the co-clustering of multivariate functional data (see for example [1]). Furthermore, we consider the problem in which each group homogenous curves itself is governed by an unknown underlying dynamical process so that the group exhibits a segmentation property. This is the co-clustering of heterogeneous and possibly dynamical multivariate functional data. We propose the functional latent block model (FLBM) to simultaneously cluster a sample of multivariate functions into a finite set of blocks, each block being an association of a cluster over individuals and a cluster over variables. Furthermore, the homogenous set within each block is modeled with a dedicated latent process functional regression model which allows its segmentation according to an underlying dynamical process. The proposed model allows thus to fully exploit the structure of the data, compared to classical latent block clustering models for continuous non functional data, which ignore the functional structure of the data, to better address the issue of high-dimension by co-clustering the variables together with the individuals, and, to further take into account the dynamical structure of the data. The FLBM can therefore serve for simultaneous co-clustering and segmentation of high-dimensional non-stationary functions. For the model inference, we propose a variational expectation-maximization (EM) algorithm (VEM-FLBM) to monotonically maximize a variational approximation of the observed-data log-likelihood. Then, we derive a stochastic version the EM algorithm to the FLBM model (SEM-FLBM). The remainder of this paper is organized as follows. Section 2 presents the proposed FLBM model. Then, Section 3 presents the developed variational EM algorithm to the unsupervised inference of the model parameters.

2 Functional Latent Block Model (FLBM)

The aim here is to cluster a sample of multivariate functions (e.g. curves, times series, signals, etc.) into a finite number of homogeneous blocks. Let us denote by $\boldsymbol{Y} = (\boldsymbol{y}_{ij})$ the data sample matrix of n individuals defined on a set $\mathcal{I}$ and d continuous functional variables defined on a set $\mathcal{J}$. Each variable is an univariate curve $\boldsymbol{y}_{ij} = \big(y_{ij}(t_1), \ldots, y_{ij}(t_{T_{ij}})\big)$ of T_{ij} observations $y(t) \in \mathbb{R}$ issued from an underlying function at the points $(t_1, \ldots, t_{T_{ij}})$, typically a sampling time.

2.1 Conditional independence and conditional data distribution

As in latent block models for binary, categorical, or continuous Gaussian data, we adopt a conditional independence assumption. We assume that there exists a partition $\boldsymbol{Z} = (z_{ik}; i = 1, \ldots, n, k = 1, \ldots, K)$ into K clusters on $\mathcal{I}$ and a partition $\boldsymbol{W} = (w_{j\ell}; j = 1, \ldots, d, \ell = 1, \ldots, M)$ into M clusters on $\mathcal{J}$, such that the univariate functions $\boldsymbol{y}_{ij}$ are conditionally independent given $\boldsymbol{Z}$ and $\boldsymbol{W}$. The z_{ik}'s (resp. $w_{j\ell}$'s) are binary indicators of row i (resp. column j) belonging to row cluster k (resp. to column cluster ℓ). The conditional probability density function (pdf) of a curve $\boldsymbol{y}_{ij}$ given the covariate vector $\boldsymbol{x}_{ij} \in \mathbb{R}^p$ (typically a vector depending on the time t), within block $k\ell$, that is, given that the ith row belongs to cluster k and the jth column belongs to cluster ℓ, is a parametric pdf of the form $f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij}; \boldsymbol{\theta}_{k\ell})$, $\boldsymbol{\theta}_{k\ell}$ being its parameter vector.

Thus, the conditional pdf of the data $\boldsymbol{Y}$ given $\boldsymbol{Z}$ and $\boldsymbol{W}$, and the predictors $\mathbf{X}$, can be expressed as

$$f(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W},\mathbf{X};\{\boldsymbol{\theta}_{k\ell}\}) = \prod_{i,j,k,\ell}\left\{f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell})\right\}^{z_{ik}w_{j\ell}}. \tag{1}$$

Many choices are possible to model the block conditional pdf in this context of co-clustering of functional data, which are here assumed to further exhibit an underlying dynamical structure, in addition to the grouping aspect and the functional representation of the data. We propose to model this distribution by a regression model, and hence the modeling of the conditional pdf in this FDA context is based on exploring the relationship of the observed variable Y_{ij} given the covariate vector $\boldsymbol{X}$ via a regression model mean function of the form $f(y|\boldsymbol{x})$, rather than only exploring the unconditional distribution of Y, as in standard latent block clustering of non-functional data. We assume that these observations, within each block kl, arise from an underling parametric possibly non-linear regression function $\mu(x(t);\boldsymbol{\beta})$ parametrized by $\boldsymbol{\beta}$. The observations of each univariate curve have a functional structure of the form: $y_{ij}(t) = \mu(x_{ij}(t);\boldsymbol{\beta}) + \epsilon(t)$ where $\epsilon(.)$ is a standard Gaussian variable representing and additive noise. Furthermore, here it is also assumed that these functional observations are governed by a hidden process with a finite number of regimes or states so that they exhibit a segmentation property in the temporal (or longitudinal) dimension.

To accommodate this further behavior for these functional data, we propose to model the conditional data distribution with a particular dynamical regression model with a hidden process, which has been shown to have more attractive properties compared to several regression models for functional data. This is the regression model with hidden logistic process (RHLP) [3, 4, 24]. The RHLP as a model for each conditional block distribution is particularly suited to approximate non-linear functions and to provide a partition of each block into a finite number of segments.

2.2 Modeling with the Regression model with hidden logistic process

In the proposed functional latent block model, each block (1) is a regression model with a hidden logistic process (RHLP). The RHLP for a block kl assumes that the observed function (a curve or a time series) $\boldsymbol{y}_{ij}$ is governed by an R-state hidden process $\boldsymbol{H} = (h_1, \ldots, h_{T_{ij}})$ with the categorical random variable $h_{ij} \in \{1, \ldots, R\}$ representing the unknown (hidden) label of the state of the observation y_{ij} at time t, where the conditional state distribution is the one of a Gaussian polynomial regression model. The polynomial regressors are therefore governed by the latent categorical variable h whose distribution is assumed to be a multinomial logistic that depends on t, and thus allows to smoothly switch from one regression model to another at each point t. The conditional distribution for each block $k\ell$ of a curve is thus defined by (see for example [3][4] for more details):

$$f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell}) = \prod_{t=1}^{T_{ij}}\sum_{r=1}^{S_{k\ell}} \alpha_{k\ell r}(t;\boldsymbol{\xi}_k)\mathcal{N}\big(y_{ij}(t);\boldsymbol{\beta}_{kr}^T\boldsymbol{x}_j(t),\sigma_{k\ell r}^2\big) \tag{2}$$

where the dynamical weights α's are given by the multinomial logistic: $\alpha_{k\ell r}(t;\boldsymbol{\xi}_k) = \frac{\exp(\xi_{k\ell r0}+\xi_{kr\ell 1}t)}{\sum_{r'=1}^{S_{k\ell}}\exp(\xi_{k\ell r'0}+\xi_{k\ell r'1}t)}$. This modeling allows to control the state transition points as well as their smoothness, and to segment each function into $S_{k\ell}$ segments by maximizing the logistic weights. The parameter vector of the conditional block distribution for the FLBM is $\boldsymbol{\theta}_{kl} = (\boldsymbol{\xi}_{k\ell}^T, \boldsymbol{\beta}_{k\ell 1}^T, \ldots, \boldsymbol{\beta}_{k\ell S_{kl}}^T, \sigma_{k\ell 1}^2, \ldots, \sigma_{k\ell S_{k\ell}}^2)^T$ where $\boldsymbol{\xi}_{k\ell} = (\boldsymbol{\xi}_{k\ell 1}^T, \ldots, \boldsymbol{\xi}_{k\ell S_{k\ell}-1}^T)^T$, $\boldsymbol{\xi}_{k\ell r} = (\xi_{k\ell r0}, \xi_{k\ell r1})^T$ being the 2-dimensional coefficient vector for the rth logistic component with $\boldsymbol{\xi}_{k\ell S_{k\ell}}$ being the null vector.

2.3 The functional latent block model

The proposed functional latent block model (FLBM) assumes, as in standard latent block models, that the label indicators z_{ij} and w_{ij} defining the partition $\boldsymbol{Z}$ over the individuals, and respectively the partition $\boldsymbol{W}$ over the attributes, are hidden. The data pdf is then given by

$$f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Psi}) \;=\; \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z},\boldsymbol{W})f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta}) \tag{3}$$

where $\mathcal{Z}$ and $\mathcal{W}$ denote the sets of possible labels z for $\mathcal{I}$ and w for $\mathcal{J}$. Moreover, we also assume that the latent variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ are independent, that is, $\mathbb{P}(\boldsymbol{Z},\boldsymbol{W})=\mathbb{P}(\boldsymbol{Z})\mathbb{P}(\boldsymbol{W})$ and are independently and identically distributed according to the multinomial distribution: $\mathbb{P}(\boldsymbol{Z})=\prod_i \mathbb{P}(z_i)$ and $\mathbb{P}(\boldsymbol{W})=\prod_j \mathbb{P}(w_j)$ with $z_i\sim\mathcal{M}(\pi_1,\ldots,\pi_K)$ and $w_j\sim\mathcal{M}(\rho_1,\ldots,\rho_M)$, where $(\pi_k=\mathbb{P}(z_{ik}=1), k=1,\ldots,K)$ and $(\rho_\ell=\mathbb{P}(w_{j\ell}=1), \ell=1,\ldots,M)$ are the non-negative mixing proportions which sum to 1. An RHLP is used as a conditional block distribution $f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta})$ given by (1). We thus obtain the following functional latent block model:

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Psi}) &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \mathbb{P}(\boldsymbol{Z};\boldsymbol{\pi})\mathbb{P}(\boldsymbol{W};\boldsymbol{\rho}) f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{W};\boldsymbol{\theta}) \\ &= \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k}\pi_k^{z_{ik}} \prod_{j,\ell}\rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell})^{z_{ik}w_{j\ell}}. \end{aligned} \tag{4}$$

where $f(\boldsymbol{y}_{ij}|\boldsymbol{x}_{ij};\boldsymbol{\theta}_{k\ell})$ is defined by (2). The model is parametrized by the parameter vector $\boldsymbol{\Psi}=(\boldsymbol{\pi}^T,\boldsymbol{\rho}^T,\boldsymbol{\theta}^T)^T$, with $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_K)^T$, $\boldsymbol{\rho}=(\rho_1,\ldots,\rho_M)^T$, and $\boldsymbol{\theta}=(\boldsymbol{\theta}_{11}^T,\ldots,\boldsymbol{\theta}_{k\ell}^T,\ldots,\boldsymbol{\theta}_{KM}^T)^T$. The proposed FLBM model can be represented by the following generative process:

$$\begin{aligned} z &\sim \text{Multinomial}(\pi_1,\ldots,\pi_K) \\ w &\sim \text{Multinomial}(\rho_1,\ldots,\rho_M) \\ \boldsymbol{y}|\boldsymbol{x},z,w &\sim f(.|\boldsymbol{x};\boldsymbol{\theta}_{z,w}) \end{aligned}$$

The next section is dedicated to the parameter estimation. We first propose a maximum-likelihood estimation via a variational EM algorithm and then a MCMC sampling via a stochastic EM extension.

3 Parameter estimation by a variational EM algorithm

In this first approach, the unknown parameter vector $\boldsymbol{\Psi}$ is estimated from an independent sample of unlabeled curves $((\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n))$ by monotonically maximizing the observed-data log-likelihood $\log L(\boldsymbol{\Psi})=\log f(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Psi})$. As in classical mixture-model based clustering, this log-likelihood can not be maximized in a closed form. The usual tool in such a context is the EM algorithm [5][20]. In order to derive the EM algorithm for the FLBM model, we need to define the log-likelihood of $\boldsymbol{\Psi}$ given the complete-data, which are composed of the observed data, the hidden cluster labels $\boldsymbol{Z}$ and $\boldsymbol{W}$, and the hidden processes $\{\mathbf{H}\}$ governing each block of the data. This is the FLBM complete-data log-likelihood, given by:

$$\begin{aligned} L_c(\boldsymbol{\Psi}) &= f(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{W},\mathbf{H}|\boldsymbol{X};\boldsymbol{\Psi}) \\ &= \sum_{i,k} z_{ik}\log\pi_k + \sum_{j,\ell} w_{j\ell}\log\rho_\ell + \sum_{i,j,k,\ell,t,r} z_{ik}w_{j\ell}h_{tr}\log\left[\alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell})\mathcal{N}\left(y_{ij}(t);\boldsymbol{\beta}_{k\ell r}^T\boldsymbol{x}_{ij}(t),\sigma_{k\ell r}^2\right)\right] \end{aligned} \tag{5}$$

where $(h_{tr}; t=1,\ldots,T_{ij}, r=1,\ldots,S_{k\ell})$ is a binary variable indicating from which state each observation $y_{ij}(t)$ within the block cluster $k\ell$ is originated at time t. The E-Step of the standard EM algorithm requires namely the calculation of the posterior joint distribution $\mathbb{P}(z_{ik}w_{j\ell}=1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})$ of the missing labels z and w due to the double missing structure over rows and over columns, which dos not factorize due to the conditional dependence on the observed curves of the row and the column labels. To tackle this problem, [11] [12] proposed a variational approximation-based solution for latent block clustering, by relying on the Neal and Hinton interpretation of the EM algorithm [22]. We adopt this variational approximation in this context of model-based co-clustering of multivariate functional data. The resulting variational EM algorithm for the FLBM (VEM-FLBM) model, starts from an initial solution at iteration $q=0$, and then alternates 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. $\tilde{z}_{ik}^{(q+1)} \propto \pi_k \exp\left(\sum_{j,\ell,t,r}\tilde{w}_{j\ell}^{(q)}\ \tilde{h}_{tr}^{(q)}\log\left[\alpha_{k\ell r}(t;\boldsymbol{\xi}_{k\ell}^{(q)})\mathcal{N}\left(y_{ij}(t);\boldsymbol{\beta}_{k\ell r}^{T^{(q)}}\boldsymbol{x}_{ij}(t),\sigma_{k\ell r}^{(q)^2}\right)\right]\right)$

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

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

where the tilde notations stand for the variational approximation with $\tilde{z}_{ik} = \mathbb{P}(z_{ik} = 1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})$, $\tilde{w}_{j\ell} = \mathbb{P}(w_{j\ell} = 1|\boldsymbol{y}_{ij},\boldsymbol{x}_{ij})$, and $\tilde{h}_{tr} = \mathbb{P}(h_{tr} = 1|z_i, w_j, y_{ij}(t), x_{ij}(t))$.

M Step update the parameters estimates $\boldsymbol{\theta}^{(q+1)}$ 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. $\rho_\ell^{(q+1)} = \frac{\sum_j \tilde{w}_{j\ell}^{(q+1)}}{d}$

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

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

   The regression parameters updates consist in analytic weighted least-squares problems:

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

5. $\sigma_{k\ell r}^{2(q+1)} = \frac{\sum_{i,j}\tilde{z}_{ik}^{(q)}\tilde{w}_{j\ell}^{(q)}\|\sqrt{\boldsymbol{\Lambda}_{ijkr}^{(q)}}(\boldsymbol{y}_{ij}-\mathbf{X}_{ij}\boldsymbol{\beta}_{kr}^{(q+1)})\|^2}{\sum_{i,j}\tilde{z}_{ik}^{(q)}\tilde{w}_{j\ell}^{(q)}\ \mathrm{trace}(\boldsymbol{\Lambda}_{ijkr}^{(q)})}$ where $\mathbf{X}_{ij}$ is the design matrix for the ith curve, $\boldsymbol{\Lambda}_{ijkr}^{(q)}$ is the diagonal matrix whose diagonal elements are the posterior segment memberships $\{\tilde{h}_{ijtr}^{(q)}; t = 1,\ldots,T_{ij}\}$.

4 Conclusion and discussion

The proposed FBLM is suited to the cluster analysis and segmentation of high-dimensional functional data arising from a population of different groups where each group is governed by an underlying hidden process. The model inference can be performed by a variational EM algorithm. This algorithm will be experimented in an ongoing work and will be applied on simulations and real-world data. The obtained results will be presented during the conference. The next problem which will be investigated is the one of model selection. In model-based co-clustering approaches, the problem of model selection in general consists in selecting the best number of blocks (co-clusters). The most commonly used penalized log-likelihood criteria such as BIC, AIC etc. can not be directly used for the block mixture models. Approximations, namely variational ones, are needed such as approximated ICL or BIC-like criteria as in [19]. In [17], the authors developed a Bayesian inference technique using MCMC for the latent block model for categorical data, and an exact ICL for model selection. These criteria might also be used for the proposed FLBM model.

References

[1] Y. Ben Slimen, S. Allio, and J. Jacques. Model-Based Co-clustering for Functional Data. HAL preprint hal-01422756, December 2016.

[2] C. Bouveyron and J. Jacques. Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Analysis and Classification*, 5(4):281–300, 2011.

[3] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. Time series modeling by a regression approach based on a latent process. *Neural Networks*, 22(5-6):593–602, 2009. PDF.

[4] F. Chamroukhi, A. Samé, G. Govaert, and P. Aknin. A hidden process regression model for functional data description. Application to curve discrimination. *Neurocomputing*, 73(7-9):1210–1221, March 2010.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society, B*, 39(1):1–38, 1977.

[6] E. Devijver. Model-based clustering for high-dimensional data. Application to functional data. Technical report, Département de Mathématiques, Université Paris-Sud, 2014.

[7] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[8] S. J. Gaffney and P. Smyth. Joint probabilistic curve clustering and alignment. In *In Advances in Neural Information Processing Systems (NIPS)*, 2004.

[9] G. Govaert and M. Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463 – 473, 2003. Biometrics.

[10] G. Govaert and M. Nadif. Fuzzy clustering to estimate the parameters of block mixture models. *Soft Computing*, 10(5):415–422, 2006.

[11] G. Govaert and M. Nadif. Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52(6):3233 –3245, 2008.

[12] G. Govaert and M. Nadif. *Co-Clustering*. Computer engineering series. Wiley-ISTE, November 2013. 256 pages.

[13] J. Jacques and C. Preda. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106, 2014.

[14] G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society Series B*, 63:533–550, 2001.

[15] G. M. James and C. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462), 2003.

[16] C. Keribin, V. Brault, G. Celeux, and G. Govaert. Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, 2012.

[17] C. Keribin, V. Brault, G. Celeux, and G. Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16, 2014.

[18] X. Liu and M.C.K. Yang. Simultaneous curve registration and clustering for functional data. *Computational Statistics and Data Analysis*, 53(4):1361–1376, 2009.

[19] A. Lomet. *Sélection de modèle pour la classification croisée de données continues*. Ph.D. thesis, Université de Technologie de Compiègne, 2012.

[20] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. New York: Wiley, second edition, 2008.

[21] G. J. McLachlan and D. Peel. *Finite mixture models*. New York: Wiley, 2000.

[22] R. Neal and G. E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants*, pages 355–368. Dordrecht: Kluwer Academic Publishers, 1998.

[23] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, June 2005.

[24] A. Samé, F. Chamroukhi, G. Govaert, and P. Aknin. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, pages 1–21, 2011.