Model-Based Co-Clustering of Multivariate Functional Data
Faicel Chamroukhi, Christophe Biernacki

To cite this version:
Faicel Chamroukhi, Christophe Biernacki. Model-Based Co-Clustering of Multivariate Functional Data. ISI 2017 - 61st World Statistics Congress, Jul 2017, Marrakech, Morocco. hal-01653782

HAL Id: hal-01653782
https://hal.archives-ouvertes.fr/hal-01653782
Submitted on 1 Dec 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Model-Based Co-Clustering of Multivariate Functional Data

Faïcel 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
Submitted on April 13, 2017

Abstract

High dimensional data clustering is an increasingly interesting topic in the statistical analysis of heterogeneous 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 homogeneous 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 heterogeneous 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 variables. The adapted paradigm of analyzing such data is the increasing framework of functional data analysis (FDA) [25]. 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 [9]), that is the one based on the flexible and statistically sound framework of mixture models [23], and the well-known desirable properties of the expectation-maximization (EM) algorithm [7, 22]. This flexible modeling framework in multivariate analysis is taking a growing investigation for FDA. See for example [2, 8, 10, 15, 16, 17, 20]. 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 [11, 13, 14], 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 [14] and have been developed for binary
Thus, the conditional pdf of the data column belongs to cluster probability density function (pdf) of a curve indicators of row $K$ in independence assumption. We assume that there exists a partition as in latent block models for binary, categorical, or continuous Gaussian data, we adopt a conditional 2.1 Conditional independence and conditional data distribution 2 Functional Latent Block Model (FLBM) the unsupervised inference of the model parameters. We propose a variational expectation-maximization (EM) algorithm (VEM-FLBM) to monotonically maximize co-clustering and segmentation of high-dimensional non-stationary functions. For the model inference, we take into account the dynamical structure of the data. The FLBM can therefore serve for simultaneous dress the issue of high-dimension by co-clustering the variables together with the individuals, and, to further models for continuous non functional data, which ignore the functional structure of the data, to better ad- model allows thus to fully exploit the structure of the data, compared to classical latent block clustering 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 $Y = \{y_{ij}\}$ 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 $y_{ij} = (y_{ij}(t_1), \ldots, y_{ij}(t_n))$ of $T_\ell$ observations $y(t) \in \mathbb{R}$ issued from an underlying function at the points $(t_1, \ldots, t_{T_\ell})$, 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 $Z = \{z_{ik}; i = 1, \ldots, n, k = 1, \ldots, K\}$ into $K$ clusters on $\mathcal{I}$ and a partition $W = \{w_{jl}; j = 1, \ldots, d, \ell = 1, \ldots, M\}$ into $M$ clusters on $\mathcal{J}$, such that the univariate functions $y_{ij}$ are conditionally independent given $Z$ and $W$. 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 $\ell$). The conditional probability density function (pdf) of a curve $y_{ij}$ given the covariate vector $x_{ij} \in \mathbb{R}^p$ (typically a vector depending on the time $t$), within block $k, \ell$, that is, given that the $i$th row belongs to cluster $k$ and the $j$th column belongs to cluster $\ell$, is a parametric pdf of the form $f(y_{ij} | x_{ij}; \theta_{k,\ell})$, $\theta_{k,\ell}$ being its parameter vector. Thus, the conditional pdf of the data $Y$ given $Z$ and $W$, and the predictors $X$, can be expressed as

$$f(Y | Z, W, X; \{\theta_{k,\ell}\}) = \prod_{i,j,k,\ell} f(y_{ij} | x_{ij}; \theta_{k,\ell})^{z_{ik}w_{j\ell}}. \quad (1)$$
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 $X$ via a regression model mean function of the form $f(y|x)$, 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 underlying parametric possibly non-linear regression function $\mu(x(t); \beta)$ parametrized by $\beta$. The observations of each univariate curve have a functional structure of the form: $y_{ij}(t) = \mu(x_{ij}(t); \beta) + \epsilon(t)$ 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) [5, 6, 26]. 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) $y_{ij}$ is governed by an $R$-state hidden process $H = (h_1, \ldots, h_{T_{ij}})$ 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 $kl$ of a curve is thus defined by (see for example [5][6] for more details):

$$f(y_{ij} | x_{ij}; \theta_{kl}) = \prod_{t=1}^{T_{ij}} \sum_{r=1}^{S_{kl}} \alpha_{ktr}(t; \xi_{kl}) N(y_{ij}(t); \beta^T_{k\ell}x_{ij}(t), \sigma^2_{k\ell r})$$

(2)

where the dynamical weights $\alpha$’s are given by the multinomial logistic: $\alpha_{ktr}(t; \xi_{kl}) = \frac{\exp(\xi_{k0r} + \xi_{ktr} t)}{\sum_{j \neq r} \exp(\xi_{k0j} + \xi_{ktr} t)}$. This modeling allows to control the state transition points as well as their smoothness, and to segment each function into $S_{kl}$ segments by maximizing the logistic weights. The parameter vector of the conditional block distribution for the FLBM is $\theta_{kl} = (\xi_{k0}^T, \beta_{k1}^T, \ldots, \beta_{kS_{kl}}^T, \sigma_{k1}^2, \ldots, \sigma_{kS_{kl}}^2)^T$ where $\xi_{kl} = (\xi_{k0}, \ldots, \xi_{kS_{kl} - 1})^T$, $\xi_{ktr} = (\xi_{k0r}, \xi_{ktr}^T)$ being the 2-dimensional coefficient vector for the $r$th logistic component with $\xi_{kS_{kl}}$ 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 $Z$ over the individuals, and respectively the partition $W$ over the attributes, are hidden. The data pdf is then given by

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

(3)

where $Z$ and $W$ denote the sets of possible labels $z$ for $I$ and $w$ for $J$. Moreover, we assume that the latent variables $Z$ and $W$ are independent, that is, $\mathbb{P}(Z, W) = \mathbb{P}(Z) \mathbb{P}(W)$ and are independently and identically distributed according to the multinomial distribution: $\mathbb{P}(Z) = \prod_i \mathbb{P}(z_i)$ and $\mathbb{P}(W) = \prod_j \mathbb{P}(w_j)$ 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 $(\rho_\ell = \ldots$
\( \mathbb{P}(w_{ij} = 1), \ell = 1, \ldots, M \) are the mixing proportions which sum to 1. An RHLP is used as a conditional block distribution \( f(Y|X, Z, W; \theta) \) (1). We thus obtain the following functional latent block model:

\[
\begin{align*}
    f(Y|X; \Psi) &= \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \mathbb{P}(Z; \pi) \mathbb{P}(W; \rho) f(Y|X, Z, W; \theta) \\
    &= \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,k} \pi_k \prod_{j,\ell} \rho_{ij}^{w_{ij}} \prod_{i,j,k,\ell} f(y_{ij}|x_{ij}; \theta_{kl})^{z_{ik}w_{ij}}.
\end{align*}
\]

where \( f(y_{ij}|x_{ij}; \theta_{kl}) \) is defined by (2). The model is parametrized by the parameter vector \( \Psi = (\pi^T, \rho^T, \theta^T)^T \), with \( \pi = (\pi_1, \ldots, \pi_K)^T \), \( \rho = (\rho_1, \ldots, \rho_M)^T \), and \( \theta = (\theta_{11}, \ldots, \theta_{kk}, \ldots, \theta_{KM})^T \). The proposed FLBM model can be represented by the following generative process:

\[
\begin{align*}
    z &\sim \text{Multinomial}(\pi_1, \ldots, \pi_K) \\
    w &\sim \text{Multinomial}(\rho_1, \ldots, \rho_M) \\
    y|x, z, w &\sim f(.|x; \theta_{z,w})
\end{align*}
\]

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 \( \Psi \) is estimated from an independent sample of unlabeled curves \((x_1, y_1), \ldots, (x_n, y_n)\) by monotonically maximizing the observed-data log-likelihood \( \log L(\Psi) = \log f(Y|X; \Psi) \). 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 [7][22]. In order to derive the EM algorithm for the FLBM model, we need to define the log-likelihood of \( \Psi \) given the complete-data, which are composed of the observed data, the hidden cluster labels \( Z \) and \( W \), and the hidden processes \( \{H\} \) governing each block of the data. This is the FLBM complete-data log-likelihood, given by:

\[
\begin{align*}
    \log L_c(\Psi) &= \log f(Y, Z, W, H|X; \Psi) \\
    &= \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,\ell} w_{j\ell} \log \rho_{ij} + \sum_{i,j,k,\ell,r} z_{ik}w_{j\ell}h_{tr} \log \left[ \alpha_{t\ell r}(t; z_{ik}, w_{j\ell}) \mathcal{N}(y_{ij}(t); \beta_{t\ell r}^{(y)}, x_{ij}(t), \sigma_{t\ell r}^{(y)^2}) \right]
\end{align*}
\]

where \((h_{tr}; t = 1, \ldots, T, r = 1, \ldots, S_{kl})\) is a binary variable indicating from which state each observation \( y_{ij}(t) \) within the block cluster \( kl \) is originated at time \( t \). The E-Step of the standard EM algorithm requires namely the calculation of the posterior joint distribution \( \mathbb{P}(z_{ik}w_{j\ell} = 1|y_{ij}, x_{ij}) \) of the missing labels \( z \) and \( w \) due to the double missing structure over rows and over columns, which does not factorize due to the conditional dependence on the observed curves of the row and the column labels. To tackle this problem, [13] [14] proposed a variational approximation-based solution for latent block clustering, by relying on the Neal and Hinton interpretation of the EM algorithm [24]. 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. \( z_{ik}^{(q+1)} \propto \pi_k^{(q)} \exp \left( \sum_{j,\ell,r} \tilde{w}_{j\ell}^{(q)} \tilde{h}_{tl}^{(q)} \log \left[ \alpha_{t\ell r}(t; z_{ik}^{(q)}, w_{j\ell}^{(q)}) \mathcal{N}(y_{ij}(t); \beta_{t\ell r}^{(y)}, x_{ij}(t), \sigma_{t\ell r}^{(y)^2}) \right] \right) \)
2. \( w_{j\ell}^{(q+1)} \propto \rho_{ij}^{(q)} \exp \left( \sum_{i,k,\ell,r} \tilde{z}_{ik}^{(q)} \tilde{h}_{tl}^{(q)} \log \left[ \alpha_{t\ell r}(t; z_{ik}^{(q)}, w_{j\ell}^{(q)}) \mathcal{N}(y_{ij}(t); \beta_{t\ell r}^{(y)}, x_{ij}(t), \sigma_{t\ell r}^{(y)^2}) \right] \right) \)
3. \( h_{tl}^{(q+1)} \propto \alpha_{t\ell r}^{(q)} \mathcal{N}(y_{ij}(t); \beta_{t\ell r}^{(y)}, x_{ij}(t), \sigma_{t\ell r}^{(y)^2}) \)

where the tilde notations stand for the variational approximated \( \tilde{z}_{ik} = \mathbb{P}(z_{ik} = 1|y_{ij}, x_{ij}) \), \( \tilde{w}_{j\ell} = \mathbb{P}(w_{j\ell} = 1|y_{ij}, x_{ij}) \), and we have \( \tilde{h}_{tl} = \mathbb{P}(h_{tr} = 1;z_{ik}, w_{j\ell}, y_{ij}(t), x_{ij}(t)) \).
M Step update the parameters estimates $\theta^{(q+1)}$ given the current estimated posterior memberships

1. $\pi_k^{(q+1)} = \frac{\sum j z_{i,j}^{(q)} \pi_j}{n}$
2. $\rho_{kj}^{(q+1)} = \frac{\sum j w_{i,j}^{(q)}}{d}$

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

3. $\xi_{kj}^{(new)} = \xi_{kj}^{(old)} \left[ \sum_{i,j} \frac{z_{i,j}^{(q)} w_{i,j}^{(q)} X_{ij}^{T} A_{ij.kr}^{(q)} X_{ij}^{(q)} \log \alpha_{ij} \log \xi_{kj}^{(old)} + \log \xi_{kj}^{(old)}}{\sum_{i,j} z_{i,j}^{(q)} w_{i,j}^{(q)} X_{ij}^{T} A_{ij.kr}^{(q)} X_{ij}} \right]^{-1}$
   
4. $\beta_{ktr}^{(q+1)} = \frac{\sum_{i,j} z_{i,j}^{(q)} w_{i,j}^{(q)} (X_{ij}^{T} A_{ij.kr}^{(q)} X_{ij} - \sqrt{\Lambda_{ij.kr}^{(q)}})}{\sum_{i,j} z_{i,j}^{(q)} w_{i,j}^{(q)} \text{trace}(\Lambda_{ij.kr}^{(q)})}$ where $X_{ij}$ is the design matrix for the $i$th curve, $A_{ij.kr}^{(q)}$ is the diagonal matrix whose elements are the posterior segment memberships $\{h_{ijr}^{(q)}; t = 1, \ldots, T_{ij}\}$.

3.1 The CEM and SEM algorithms for the FLBM

We also use another approximation fo the EM algorithm, that is the Classification EM (CEM) algorithm [4]. In the CEM framework, a maximum classification likelihood estimation is performed rather than a maximum likelihood estimation. This is achieved by replacing the fuzzy memberships, which correspond to the posterior component memberships in the previous scheme, by their binary allocations. The SEM algorithm [3] allows to overcome some drawbacks of the EM algorithm, including its sensitivity to starting values and does not use an approximation. It was used for for latent block models for categorical data [19] and more recently for functional data [1]. For the proposed FLBM model, the SEM algorithm steps and those the VEM are essentially the same; the SEM incorporates a stochastic step between the VE and the M steps where the missing labels are simulated according to their posterior distribution given a current estimate of the model.

Compared to [1], the proposed FLBM is adapted for both functional data approximation and to deal with dynamical functions by segmentation. Also from a perspective of model-selection, while the model in [1] uses transformed data by functional PCA, which makes the use of information criteria not adapted, and thus can not be compared, from this model selection point of view, to the the proposed FLBM model, which directly models the original data.

4 Conclusion and discussion

The proposed FLBM 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 or an SEM algorithm, which does not use approximation. 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 [21]. In [19], 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


