Model-based co-clustering for functional data

Yosra Ben Slimen, Sylvain Allio, Julien Jacques

Model-Based Co-clustering for Functional Data

Yosra Ben Slimen [1,2], Sylvain Allio [1] & Julien Jacques [2]

[1] *Orange Labs, Belfort, France*
[2] *Univ Lyon, Lumière Lyon 2, ERIC, Lyon, France*
yosra.benslimen@orange.com, sylvain.allio@orange.com, julien.jacques@univ-lyon2.fr

Résumé. Nous présentons dans ce travail un algorithme de coclustering pour données fonctionnelles. Cet algorithme repose sur le modèle des blocs latents utilisant une modélisation gaussienne des composantes principales fonctionnelles et un algorithme SEM-Gibbs pour l'inférence.

Mots-clés. coclustering, données fonctionnelles, algorithme SEM-Gibbs

Abstract. A model-based coclustering algorithm for functional data is presented. This algorithm relies on the latent block model using a Gaussian model for the functional principal components and a SEM-Gibbs algorithm for inference.

Keywords. model-based coclustering, functional data, SEM-Gibbs algorithm

1 Introduction

With the introduction of new technologies and services in mobile networks, the complexity of these latter have increasingly grown creating an heterogeneous environment where different architectures (micro-, macro-, pico-, femto-cells) and different radio access technologies (GSM, UMTS, LTE, ...) coexist. In this context, mobile operators need to deal with the new challenges in order to provide a top quality of services without increasing costs.

The quality of services is measured by data captured from the network. These data are generated from different sources such as probes, robots and key performance indicators (KPI). They are used for different needs such as network maintenance, optimization, troubleshooting and management. They are also used for self-organizing networks e.g for self-healing and self-optimization tasks.

KPIs are measurements used to monitor quality of service perceived by the user and the network performance. The KPIs concern different network elements such as transceivers, cells, sites, ... and they are defined by mathematical formulas derived from different counters, computed periodically from the network. Therefore, KPIs are continuous data that are computed with different temporal granularities (hourly or less, daily, weekly). Figure 1 provides an illustration of 20 daily evolutions of 30 KPIs, with a temporal granularity of 15 minutes.

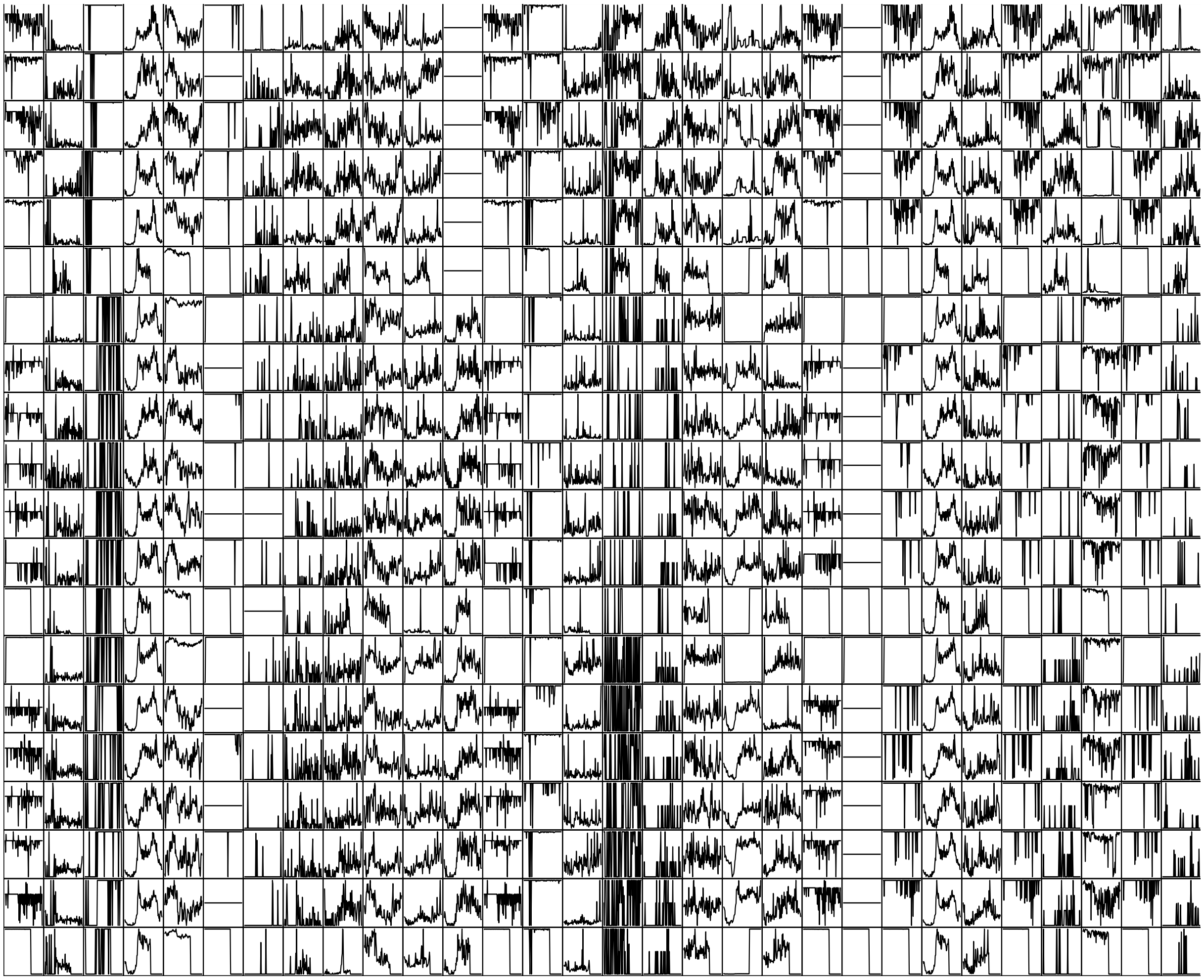

Figure 1: Sample of daily evolutions of 30 KPIs.

Some KPIs are common across the different radio access technologies while others are specific to each one of them. Even under the same technology, the counters/KPIs may differ from one constructor to another. Therefore, as the number of technologies, services, cell types, and constructors grows, the number of KPIs observed by the support team becomes enormous and network planning and operating becomes more complex. However, nowadays, the number of procedures, e.g related to optimization and fault management, that are manually carried out is still considerable. This has triggered a significant research effort, gathered under the term "automatic networking".

In this context, with this work, we are interested in KPIs and the study of their behaviors. We define a clustering of not only the different KPIs, but also of observations.

The proposed model can help the technical support team of mobile operators and it can afford additional information to self-organizing networks by discovering new relationships between the different KPIs and new similarities between their behaviors in different days. In other terms, we define a model for co-clustering of both KPIs and days.

The rest of this paper is organized as follows. Section 2 presents the data representation and its pre-processing. Section 3 presents the latent block model for functional data that we propose and Section 4 details its inference. Finally, Section 5 describes briefly the ongoing work.

2 The data

The data under study are a sample of n observations (days), each observation being described by a set of p curves (functional features, KPIs). The statistical model underlying data represented by (multivariate) curves is a stochastic process with continuous time:

$$\mathbf{X} = \{\mathbf{X}(t)\}_{t\in[0,T]} \quad \text{with} \quad \mathbf{X}(t) = (X_1(t), \ldots, X_p(t))' \in \mathbb{R}^p, \quad p \geq 2.$$

A sample path of $\mathbf{X}$ is represented by a set of p curves.

2.1 Transformation of the observed discretized curves

In practice, data are generally observed at discrete time points and with some noise. In order to reflect the functional nature of data, smoothing methods consider that the true curve belongs to a finite dimensional space spanned by some basis of functions. Let us 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 $\{\phi_{j\ell}\}_{\ell=1,\ldots,m_j}$:

$$x_{ij}(t) = \sum_{\ell=1}^{m_j} a_{ij\ell}\phi_{j\ell}(t), \quad t \in [0, T]. \tag{1}$$

The basis expansion coefficients $\mathbf{a}_{ij} = \{a_{ij\ell}\}_{\ell=1,\ldots,m_j}$ can be estimated by least square smoothing (see [3] for instance). In this work, the same basis $\{\phi_\ell\}_{\ell=1,\ldots,m}$ is used for all the functional features (KPIs).

2.2 Principal components analysis for functional data

Principal components analysis is often used to give a simple representation of the functional data. Principal components analysis for functional data (FPCA, [3]) consists in computing the principal components C^h and principal factors f^h of the Karhunen-Loeve expansion:

$$X(t) = \mu(t) + \sum_{h\geq 1} C^h f^h(t), \qquad t \in [0, T]. \tag{2}$$

When curves are assumed to be decomposed into a finite basis of function (1), FPCA consists in a classical PCA of the basis expansion coefficients using a metric defined by the inner product between the basis functions. FPCA has been extended to multivariate functional data in [2].

In theory, the number of principal components are infinite. However, in practice, due to the fact that the curves are observed at discrete time points and then approximated on a finite basis of functions, the maximum number of components one can compute is equal to the number m of basis functions used for approximation.

3 Latent block model for functional data

In a co-clustering study, the goal is to gather observations as well as (functional) features into clusters. For this, we use the latent block model which assumes that data into a block (defined by a cluster of observations and a cluster of features) are independent and identically distributed. Let K_r be the number of clusters in row and let K_c be the number of clusters in column.

Let $\mathbf{x} = (x_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ be the matrix of curves $(x_{ij} : x_{ij}(t),\ t \in [0,T])$ whose rows are observations (days) and whose columns are the functional features. Let $\mathbf{c} = (c_{ij}^h)_{1\leq i\leq n,1\leq j\leq p,1\leq h\leq m}$ be the principal components resulting from a (univariate) FPCA of all the curves $\mathbf{x}$, without distinction between curves from different observations or different features. In the following, the straightforward ranges for i, j, h, k_r and k_c will be omitted for simplicity of notations.

The latent block model for functional data we propose is defined by:

$$\mathrm{p}(\mathbf{x};\theta) = \sum_{\mathbf{v}\in V}\sum_{\mathbf{w}\in W} \mathrm{p}(\mathbf{v};\theta)\mathrm{p}(\mathbf{w};\theta) f(\mathbf{c}|\mathbf{v},\mathbf{w};\theta) \tag{3}$$

where,

- V is the set of all possible partitions of the rows into K groups, W is the set of all possible partitions of the columns into M groups,
- $\mathrm{p}(\mathbf{v};\theta) = \prod_{ik_r} \alpha_{k_r}^{v_{ik_r}}$, α_{k_r} being the row-mixing proportion and $\mathbf{v} = (v_{ik_r})_{ik_r}$ with $v_{ik_r} = 1$ if observation i belongs to the row cluster k_r, 0 otherwise,
- $\mathrm{p}(\mathbf{w};\theta) = \prod_{jk_c} \beta_{k_c}^{w_{jk_c}}$, β_{k_c} being the column-mixing proportion, and $\mathbf{w} = (w_{jk_c})_{jk_c}$ with $w_{jk_c} = 1$ if the functional covariate j belongs to the column cluster k_c, 0 otherwise,
- $f(\mathbf{c}|\mathbf{v},\mathbf{w};\theta) = \prod_{ijk_rk_c} \mathrm{p}(\mathbf{c}_{ij};\mu_{k_rk_c},\Sigma_{k_rk_c})^{v_{ik_r}w_{jk_c}}$ with,
  - $\mathbf{c}_{ij} = (c_{ij}^h)_{1\leq h\leq m}$ are the functional principal components of $x_{ij}(t)$,

  - $\mathrm{p}(\cdot;\mu_{k_rk_c},\Sigma_{k_rk_c})$ is the m-variate Gaussian density with mean $\mu_{k_rk_c}=(\mu_{hk_rk_c})_{1\leq h\leq m}$ and diagonal variance matrix $\Sigma_{k_rk_c}$ with diagonal $(\sigma^2_{hk_rk_c})_{1\leq h\leq m}$,

- $\theta=(\alpha_{k_r},\beta_{k_c},\mu_{k_rk_c},\Sigma_{k_rk_c})_{1\leq k_c\leq K_c,1\leq k_r\leq K_r}$.

4 Inference via a SEM-Gibbs algorithm

Let us assume that a FPCA has been carried out on the whole data set of curves, and then each curve x_{ij} is represented by its principal components c^h_{ij} for $h=1,\ldots,m$.

Inference of the latent block model is computationally infeasible with an EM algorithm [1] and we choose to use its stochastic version SEM coupled with a Gibbs sampling. Starting from an initial value of the parameter $\theta^{(0)}$ and of the missing data $\mathbf{w}^{(0)}$, the q^{th} iteration of the partial SEM-Gibbs alternates the following SE and M steps.

SE step Execute a small number (at least 1) of successive iterations of the two following steps:

1. generate the row partition $v^{(q+1)}_{ik_r}|\mathbf{c},\mathbf{w}^{(q)}$ for all $1\leq i\leq n, 1\leq k_r\leq K_r$:

$$\mathrm{p}(v_{ik_r}=1|\mathbf{c},\mathbf{w}^{(q)};\theta^{(q)})=\frac{\alpha^{(q)}_{k_r}f_{k_r}(\mathbf{c}_i|\mathbf{w}^{(q)};\theta^{(q)})}{\sum_{k'_r}\alpha^{(q)}_{k'_r}f_{k'_r}(\mathbf{c}_i|\mathbf{w}^{(q)};\theta^{(q)})}$$

   where $\mathbf{c}_i=(c^h_{ij})_{j,h}$ and $f_{k_r}(\mathbf{c}_i|\mathbf{w}^{(q)};\theta^{(q)})=\prod_{jk_c}\mathrm{p}(\mathbf{c}_{ij};\mu^{(q)}_{k_rk_c},\Sigma^{(q)}_{k_rk_c})^{w^{(q)}_{jk_c}}$

2. generate the column partition $w^{(q+1)}_{jk_c}|\mathbf{c},\mathbf{v}^{(q+1)}$ for all $1\leq j\leq p, 1\leq k_c\leq K_c$:

$$\mathrm{p}(w_{jk_c}=1|\mathbf{c},\mathbf{v}^{(q+1)};\theta^{(q)})=\frac{\beta^{(q)}_{k_c}f_{k_c}(\mathbf{c}_j|\mathbf{v}^{(q+1)};\theta^{(q)})}{\sum_{k'_c}\beta^{(q)}_{k'_c}f_{k'_c}(\mathbf{c}_j|\mathbf{v}^{(q+1)};\theta^{(q)})}$$

   where $\mathbf{c}_j=(c^h_{ij})_{i,h}$ and $f_{k_c}(\mathbf{c}_j|\mathbf{v}^{(q+1)};\theta^{(q)})=\prod_{ik_r}\mathrm{p}(\mathbf{c}_{ij};\mu^{(q)}_{k_rk_c},\Sigma^{(q)}_{k_rk_c})^{v^{(q+1)}_{ik_r}}$

M step Estimate $\theta^{(q+1)}$ conditionally on $\mathbf{v}^{(q+1)},\mathbf{w}^{(q+1)}$:

$$\begin{aligned}
\alpha^{(q+1)}_{k_r} &= \frac{1}{n}\sum_i v^{(q+1)}_{ik_r} \qquad \beta^{(q+1)}_{k_c}=\frac{1}{p}\sum_j w^{(q+1)}_{jk_c}\\
\mu^{(q+1)}_{k_rk_c} &= \frac{1}{n^{(q+1)}_{k_rk_c}}\sum_i\sum_j \mathbf{c}_{ij}^{v^{(q+1)}_{ik_r}w^{(q+1)}_{jk_c}}\\
{\Sigma_{k_rk_c}}^{(q+1)} &= \frac{1}{n^{(q+1)}_{k_rk_c}-1}\sum_i\sum_j(\mathbf{c}_{ij}-\mu^{(q+1)}_{k_rk_c})^t(\mathbf{c}_{ij}-\mu^{(q+1)}_{k_rk_c})^{v^{(q+1)}_{ik_r}w^{(q+1)}_{jk_c}}
\end{aligned}$$

where $n^{(q+1)}_{k_rk_c}=\sum_i\sum_j v^{(q+1)}_{ik_r}w^{(q+1)}_{jk_c}$.

Choosing the parameter estimation and the final partition After a brun in period, the final estimation $\hat{\theta}$ of the parameter θ is defined by the mean of sample distribution. The final partition is estimated by maximum a posteriori according to $\hat{\theta}$.

5 Ongoing work

In this paper, we defined a co-clustering model for functional data. Our ongoing work includes the establishment of an experimental study in order to test our approach. The experimentation is composed of tests with simulated data as well as tests with real data. The results will be presented during the conference.

Our model is characterized to be generic so it can be applied to any application that uses multivariate functional data and that aims to discover the relationship between them and the different observations. In particular, in the mobile network domain, our model can extract new knowledge about KPIs which will help the technical support team of network operators in optimization and troubleshooting tasks. Moreover, it can be applied as a pre-processing step of self-organizing networks procedures in order to explore the enormous number of the used KPIs.

The main advantage of this model is that it offers a way to get labeled data, which is very hard to obtain in the area of mobile network. Actually, in some use cases such as self-healing, where there is no dataset available (neither issued from live networks nor artificial), designing and evaluating data mining approaches is still a challenge that should be tackled. Therefore, starting from unlabeled data, we proposed a technique of clustering which result i.e the different clusters can be addressed to mobile network experts who can identify them. The resulted labeled data will allow the use of supervised techniques and prediction of new observations.

References

[1] G. Govaert and M. Nadif. *Co-Clustering.* Wiley-ISTE, 2013.

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

[3] J. O. Ramsay and B. W. Silverman. *Functional data analysis.* Springer Series in Statistics. Springer, New York, second edition, 2005.