Bayesian lasso functional clustering for time-course and longitudinal data

Folly Adjogou[1], Alejandro Murua[1*] and Wolfgang Raffelsberger[2]

[1] Département de mathématiques et de statistique, Université de Montréal,
CP 6128, succ. centre-ville, Montréal, Québec H3C 3J7 Canada

[2] I.G.B.M.C. Institut de Génétique et de Biologie Moléculaire et Cellulaire,
1 rue Laurent Fries, 67404 Illkirch Strasbourg, France

Abstract

The underlying molecular mechanisms triggered by critical exposure to tobacco smoke are investigated using time-course microarray transcriptomes of rats exposed to cigarette smoke. We propose a Bayesian longitudinal data analysis framework that combines functional model-based clustering and the lasso and elastic-net penalization to identify groups of genes that share similar time-evolving patterns. Individual time-course or longitudinal curves are assumed to be spanned by a finite basis of orthogonal functions. The space of coefficients of the curve expansions form an embedding space of unknwon dimension. Clusters of curves are searched for in this space, whose elements are assumed to be generated by a mixture of Student's t-distributions. The lasso penalty is used to elucidate the intrinsic dimension of the data, as well as to find an appropriate number of functional data clusters. The proposed model is used to shed light into the molecular dynamics involved in chronic obstructive pulmonary disease.

Key words: Functional data analysis, gene expression, mixture model, model-based clustering, penalized likelihood.

1 Introduction

Exposition to tobacco smoke at long-term chronic levels as well as at acute high levels represents known risks to human health. In order to understand the initial molecular events of chronic obstructive pulmonary disease (COPD) that leads to smoking related symptoms, Stevenson et al. [2007] studied the microarray transcriptome of rats exposed to cigarette smoke.

[*] Corresponding author: A. Murua. E-mail: murua@dms.umontreal.ca

During a period of 34 weeks, male Spague-Dawley rats were examined in a time-course study. This consisted of triplicate measures at 12 precise time points for both groups of rats, exposed and not exposed to tabacco smoke. The interval periods were chosen so that time-points 2 to 5 may be considered as early stage exposure with acute symptoms, while time-points 8 to 13 may be considered as prolonged exposure with chronic symptoms. The initial study analyzed the data with t-statistics associated with gene expression differences between exposed and control rats. These revealed a strong presence of upregulation of metabolic processes accompanied by stress response and genes involved in inflammation. During the later phase of smoke exposure the expression of genes related with immunity, and defense progressively increased.

We are interested in shedding further light into the mechanisms underlying critical exposure to tobacco smoke by extending the analysis of these data. We propose using a hybrid approach that marries functional data analysis with model based clustering. The idea is to find groups of genes or proteins that are either upregulated or downregulated at the different stages of the exposure. Previous analysis of gene expression profiles were not guided by the temporal profile of genes. Instead, they isolated a few genes based on individual t-tests on selected days in the acute or chronic phase of smoke exposure. With these few selected genes, it seemed possible to separate genes involved in the acute phase from genes involved in the chronic phase. Here, we present a different approach that considers all time points. Our methodology does not biased the analysis to the most extreme changes as in the initial previous study. The functional approach is needed since the data consist of time-course expressions; the model-based approach is used to cluster the time-courses. This is roughly done by reducing the dimension of the functional data to a set of latent variables which are in turn used for clustering. The discovery of the latent variables and the clustering are done simultaneously in the proposed model using an elastic-net type penalization approach embedded into a Bayesian framework.

Longitudinal data are usually analyzed using linear mixed models (Laird and Ware [1982]; Verbeke and Molenberghs [2000]). These models explicitly decompose the variation in the data into between and within-subject variability. The work of Zhao et al. [2004] has shown that functional data analysis can be a very useful complementary technique. Functional data analysis which was primarily designed for the analysis of random trajectories and infinite-dimensional data, is rapidly evolving. Many interesting procedures incorporating this approach have recently emerged in statistics and bioinformatics so as to analyze time-course gene expression data. Clustering and classification techniques are two of the major applications of the functional approach with this type of genomic data (Ullah and Finch [2013]). By definition, functional data clustering is used to search for natural groupings of data with similar characteristics. Recently, Jacques and Preda [2014] reviewed the main literature on functional data clustering. They noted that most approaches fall within three broad categories: (a) a two-stage method consisting of applying dimension reduction techniques to the data before performing clustering; (b) a machine learning approach that uses nonparametric techniques on specifically defined distances or dissimilarities between curves; and (c) a model-based clustering approach which assumes a probabilistic mixture distribution on either the principal components (the FPCA *scores*) or the *expansion coefficients* associated with a functional data expansion into a finite dimensional basis of functions. Our present work falls into this latter category.

James and Sugar [2003] seems to have been the first authors to introduce a functional model-based clustering method. Their functional model incorporates a Gaussian-mixture to describe the expansion coefficients associated with a finite spline basis. For rougher curves, Giacofci et al. [2013] proposed a Gaussian-mixture model on a wavelet decomposition of the curves. A different approach has been proposed by Same et al. [2011]. These authors assume that the curves arise from a mixture of regressions on a polynomial basis, with possible changes in regime at each instant of time.

We introduced here a Bayesian model based on splines in which the clusters are modeled by a mixture of Student's t-distributions. The multivariate student-t distribution allows to consider fatter tails than those of the normal distribution. In our experiments, we note that time-course gene-expression data (the kind of data we are dealing with in this paper) are better modeled by t-mixtures than Gaussian ones. The thickness of the tails depends on the number of degrees of freedom which can be chosen as a parameter to be estimated. Our method is useful for the analysis and clustering of general complete or sparse time-course or longitudinal data. It is inspired by recent works in variable selection for clustering of high-dimensional data (see for example Bouveyron and Brunet [2014] for a nice review). Nowadays, penalizing criteria for clustering are the preferred methods for high-dimensional variable selection. Since the pioneering work of Tibshirani [2011], where the lasso was introduced, several works on model based clustering have introduced L_1 or L_∞ penalty terms in the log-likelihood function (Pan and Shen [2007], Wang and Zhou [2008]). This is done to yield model sparsity in the form of variable selection (which may also be seen as a form of dimension reduction). Traditionally, the lasso penalizes the absolute values (L_1-norm) of coefficients that are key to the model. Our procedure uses a double lasso-penalty in the clustering criterion in order to yield optimal choices for the reduced dimension of the data (similar to variable selection in the regression context) and the number of clusters. The strength of the lasso regularization is then determined by two penalization parameters whose optimal values are unknown. Usually, these parameters are *tuning* parameters, that is, the model is estimated for some particular values of these parameters. Their optimal values are usually determined by cross-validation techniques. Note that for large datasets, cross-validation may be computationally very costly. In this work, we argue instead for a Bayesian method to elucidate the penalization parameters. Cross-validation is not needed. The regularization parameters are incorporated in the model through a lasso-driven prior distribution on the cluster mean vectors.

The paper is organized as follows. Section 2 introduces the model-based clustering with lasso penalty (model, parameter estimation, implementation). Section 3 discusses the model selection method. Section 4 describes a simulation study and comparison with existing methods. In Section 5 we apply our methodology to the analysis of the tobacco exposure data.

2 Bayesian functional clustering with lasso penalization

In this section we introduce a Bayesian framework for a penalized functional model-based clustering method. The model combines functional principal components analysis and model-based clustering in a mixed effects model. Estimation is performed through the expectation-maximization (EM) algorithm (Dempster et al. [1977]).

2.1 The unpenalized functional clustering model

We recall below the main characteristics of the basis model. If $Y_i(\cdot)$ denotes the source function that originally generates the n_i observed measurements $\mathbf{Y}_i = (y_{i1}, ..., y_{in_i})$ at time points $\mathbf{t}_i = (t_{i1}, ..., t_{in_i})$ for the individual i in a longitudinal study, it's evaluation at a specific time t is assumed to be decomposed in the form:

$$\begin{cases} Y_i(t) = \hat{\mu}(t) + \hat{\boldsymbol{f}}(t)^\mathsf{T}\boldsymbol{\alpha}_i + \epsilon_i(t) \quad (i = 1, ..., N) \\ \boldsymbol{\alpha}_i = \boldsymbol{\mu}_{\mathbf{z}_i} + \boldsymbol{\gamma}_i^{\mathbf{z}_i} \end{cases} \tag{1}$$

where $\hat{\mu}(t)$ is an overall mean function; the functions $\hat{f}_j(t)$ are the k functional principal components (FPC) with $\hat{\boldsymbol{f}}(t)^\mathsf{T} = (\hat{f}_1(t), \ldots, \hat{f}_k(t))$ and the $\epsilon_i(t)$ are error terms. The k-dimensional vectors $\boldsymbol{\alpha}_i = (\alpha_{i1}, ..., \alpha_{ik})$ are the component scores representing the coefficients of $Y_i(t)$ on the FPCs. Furthermore, a mixed effects framework is imposed on the clustering model through the component scores. Indeed, the component scores of each individual i are expressed as the sum of the individual i's cluster mean $\boldsymbol{\mu}_{\mathbf{z}_i}$ and his own effect $\boldsymbol{\gamma}_i^{\mathbf{z}_i}$ (or the deviation from it's cluster effect). The variables $\mathbf{z}_i$ are the cluster membership indicators. These are in general unknown, and clustering analysis consists on their estimation. Thus, the combination of the two expressions of equation (1) yields a 3-term decomposition for the curve $Y_i(t)$ in addition to the error term $\epsilon_i(t)$: the overall mean $[\hat{\mu}(t)]$, the cluster effect $\left[\hat{\boldsymbol{f}}(t)^\mathsf{T}\boldsymbol{\mu}_{\mathbf{z}_i}\right]$ and the individual-specific effect $\left[\hat{\boldsymbol{f}}(t)^\mathsf{T}\boldsymbol{\gamma}_i^{\mathbf{z}_i}\right]$. All those terms are rewritten in a matrix form using the specification of the model in a finite-dimensional basis $\mathbf{b}(t)^\mathsf{T} = (b_1(t), \ldots, b_q(t))$ of B-splines to obtain the following expression:

$$\mathbf{Y}_i = \mathbf{B}_i\boldsymbol{\theta}_\mu + \mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\gamma}_i^{\mathbf{z}_i} + \boldsymbol{\epsilon}_i. \tag{2}$$

In equation (2), the n_i-dimensional vector $\mathbf{Y}_i$ contains the observed measurements at time points $\mathbf{t}_i$ and $\mathbf{B}_i = [\mathbf{b}(t_{i1}), ..., \mathbf{b}(t_{in_i})]^\mathsf{T}$ is the matrix of the spline basis evaluated at those time points. The q-dimensional vector $\boldsymbol{\theta}_\mu$ and the matrix $\boldsymbol{\Theta}$ represent, respectively, the coefficients in the basis of the overall mean function $\hat{\mu}(t) = \mathbf{b}(t)^\mathsf{T}\boldsymbol{\theta}_\mu$ and the principal components functions $\hat{\boldsymbol{f}}(t)^\mathsf{T} = \mathbf{b}(t)^\mathsf{T}\boldsymbol{\Theta}$. The measurement errors $\boldsymbol{\epsilon}_i$ are assumed to follow a multivariate Student's t distribution with unknown degrees of freedom ν_0. The functional model-based clustering in equation (2) is embedded into a Bayesian framework and the following assumptions are made to complete the setup.

$$\begin{cases} \mathbf{z}_i \sim \text{Multinomial}(1; \pi_1, \dots, \pi_G) \quad \text{with} \quad (\pi_1, \dots, \pi_G) \sim \text{Dirichlet}(a_1, \dots, a_G) \\ \boldsymbol{\mu}_g \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Gamma}_\mu) \quad \text{and} \quad \boldsymbol{\Gamma}_\mu \sim \text{InvWishart}(m, (m-k-1)I_k) \\ \boldsymbol{\gamma}_i^g \sim \mathcal{N}_k(\mathbf{0}, \boldsymbol{\Gamma}_g) \quad \text{and} \quad \boldsymbol{\Gamma}_g \sim \text{InvWishart}(m, (m-k-1)\mathbf{D}) \\ \mathbf{D} = \text{diag}(d_{11}, d_{22}, \dots, d_{kk}) \quad \text{with} \quad d_{jj} \sim \text{Inv}\,\chi^2(m) \quad \text{and i.i.d} \quad (j = 1, \dots, k) \\ \left[\boldsymbol{\epsilon}_i | \nu_i \sim \mathcal{N}_{n_i}(0, \sigma^2 \nu_i I_{n_i}) \quad \text{and} \quad \nu_i \sim \text{Inv}\,\chi^2(\nu_0)\right] \quad \Rightarrow \quad \left[\boldsymbol{\epsilon}_i \sim t_{\nu_0}(0, \sigma^2 I_{n_i})\right] \\ \text{with} \quad \sigma^2 \sim \text{InvGamma}(\alpha_\sigma, \beta_\sigma). \end{cases} \tag{3}$$

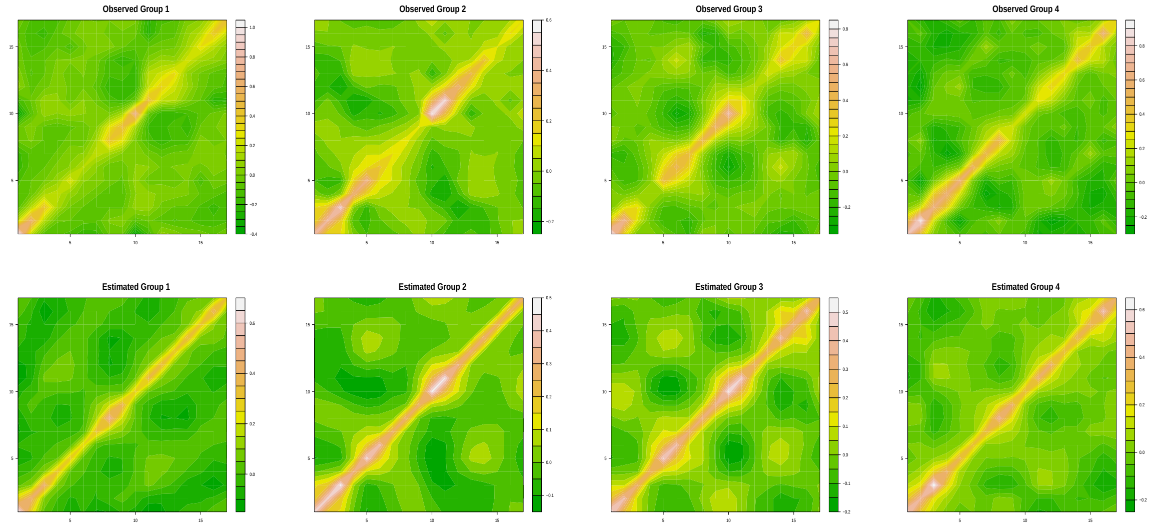
In general, the clustering of the individuals based on this model relies essentially on a space of much reduced dimension than the original longitudinal trajectories through the decomposition on the functional principal components. Note that choosing $k \leq 3$ would allow some form of visualization of the groups. The model of James and Sugar [2003] add another layer of parametrization in the clusters so that the clustering is forced to lie in a subspace of very small dimension. We prefer to let the data tell us which dimension better describe the clustering structure.

Note that the model suppose that the variance-covariance matrix of the vector of measurements Y is equal to:

$$\begin{aligned} Var(Y) &= Var(E(Y|g)) + E(Var(Y|g)) \\ &= Var(\boldsymbol{B\theta_\mu} + \boldsymbol{B\Theta\mu_g}) + E(\boldsymbol{B\Theta\Gamma_g\Theta^T B^T} + \sigma^2 I) \\ &= \boldsymbol{B\Theta}(\boldsymbol{\Gamma_\mu} + \boldsymbol{D})\boldsymbol{\Theta^T B^T} + \sigma^2 I. \end{aligned}$$

This can be estimated using the Maximum A Posteriori (MAP) estimators of the parameters, which are in turn obtained by the EM algorithm described in the Appendix A. As an illustration of this covariance estimate, consider the following yeast cycle data consisting of the fluctuations of the expression levels of about 6000 genes over two cell cycles comprising 17 time points. For purposes of this example, we consider the 5-phase subset of the data Cho et al. [1998]. It consists of 386 genes which have been assigned to one of the five phases of the cells cycle. Experts in yeast cell cycle estimated that five phases are present during the yeast cell cycle. A Bayesian information criterion (BIC) based selection procedure applied with our model yields only four clusters. Figure 1 shows the observed and estimated variance-covariance matrices associated with the four clusters found by our procedure.

Figure 1: Yeast cycle data. The observed (top row) and estimated (lower row) variance-covariance matrices by cluster. The clusters are arranged from left to right, starting with Cluster 1.



In the perspective of the EM algorithm used to estimate the parameters, the log-likelihood of the model as stated in equations (2) and (3) is obtained by considering the « complete data » $(\mathbf{Y}, \mathbf{W})$ where $\mathbf{Y} = \{\mathbf{Y}_1, ..., \mathbf{Y}_N\}$ denotes the observed data composed of the N longitudinal trajectories $(i = 1, ..., N)$ and $\mathbf{W} = \{\mathbf{z}_1, ..., \mathbf{z}_N, \boldsymbol{\gamma}_1^{\mathbf{z}_1}, ..., \boldsymbol{\gamma}_N^{\mathbf{z}_N}\} = \{\vec{\mathbf{z}}, \vec{\boldsymbol{\gamma}}^{\vec{\mathbf{z}}}\}$ denotes the missing data composed of the cluster indicators and the individual-specific effects. Let $\boldsymbol{\Pi}$ denote the set of the model parameters to be estimated and, let $\mathfrak{L}(\mathbf{Y}, \mathbf{W}; \boldsymbol{\Pi}) = \log[p(\mathbf{Y}, \mathbf{W}; \boldsymbol{\Pi})]$ denote the log-likelihood derived from the distributions involved in the model. Note that $\boldsymbol{\Pi} = \{\vec{\nu}, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}\}$ where

$$\begin{cases} \vec{\nu} = \{\nu_1, ..., \nu_N\}; \quad \vec{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_1, ...\boldsymbol{\mu}_G\}; \quad \vec{\boldsymbol{\Gamma}} = \{\boldsymbol{\Gamma}_1, ..., \boldsymbol{\Gamma}_G\}; \\ \boldsymbol{\Lambda} = \{\boldsymbol{\theta}_\mu, \boldsymbol{\Theta}, \mathbf{D}, \boldsymbol{\Gamma}_\mu, \pi_1, \pi_2, ..., \pi_G, \nu_0, \sigma^2\}. \end{cases} \tag{4}$$

The expression of $\mathfrak{L}(\mathbf{Y}, \mathbf{W}; \boldsymbol{\Pi})$ is presented in Appendix A as well as the details leading to its computation.

2.2 The penalized log-likelihood

One of the main goals of our methodology is to adequately determine the two characteristic model parameters: the number of clusters G and the dimension k of the functional principal components $\hat{\boldsymbol{f}}(t)$. The dimension q of the B-splines basis is not considered as a parameter. The number of basis functions is either set to a specific value with respect to the measurement time points of all individuals, or indirectly defined by supplying the break points or knots. The motivation for this decision comes from extensive simulations [Adjogou, 2017] which indicates that the value of the parameter q has very little influence on the clustering results as measured by the Adjusted Rand Index (ARI) scores [Rand, 1971, Hubert and Arabie, 1985].

In this new framework, we choose to estimate k and G by penalizing the log-likelihood function. The two penalizations are lasso-type ones. The main objective is to obtain a sparse solution with many estimates of cluster means basis coefficients automatically shrinked, thus realizing dimension reduction and with many inter-cluster distances shrinked, thus merging homogeneous clusters. The proposed penalized log-likelihood function is defined as

$$\begin{aligned}\mathfrak{L}^{pen}(\mathbf{Y},\mathbf{W};\mathbf{\Pi}) &= \mathfrak{L}(\mathbf{Y},\mathbf{W};\mathbf{\Pi}) - P_\lambda(\mathbf{\Pi}) \qquad (5)\\ &= \log[p(\mathbf{Y},\mathbf{W};\mathbf{\Pi})] - P_\lambda(\vec{\boldsymbol{\mu}})\end{aligned}$$

where $\vec{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_1, ...\boldsymbol{\mu}_G\}$ and $P_\lambda(\cdot)$ denotes a lasso-type penalty function with tuning parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. A lasso-type penalty function is applied to (i) the cluster mean components associated with each dimension, and to (ii) the distances between cluster means. The penalization $P_\lambda(\vec{\boldsymbol{\mu}})$ takes the form:

$$\begin{aligned}P_\lambda(\vec{\boldsymbol{\mu}}) &= \sum_{j=1}^{k} P_{\lambda_1}\left(\sum_{g=1}^{G} |\mu_g^j|\right) + \sum_{g=1}^{G}\sum_{h=1}^{g} P_{\lambda_2}\left(D^{ist}(\boldsymbol{\mu}_g, \boldsymbol{\mu}_h)\right)\\ &= \lambda_1 \sum_{j=1}^{k}\left(\sum_{g=1}^{G} |\mu_g^j|\right) + \lambda_2 \sum_{g=1}^{G}\sum_{h=1}^{g} D^{ist}(\boldsymbol{\mu}_g, \boldsymbol{\mu}_h) \qquad (6)\end{aligned}$$

where the function $D^{ist}(.,.)$ can optionally be the L_1 norm distance or the L_2 norm distance. The first term of the penalty function which is associated with the hyperparameter λ_1 is used to shrink towards zero, for a given j, the estimates $(\sum_{g=1}^{G} |\mu_g^j|)$ which are close to zero. As a consequence, the model will reduce the dimension k of the cluster space by eliminating those principal components j that are irrelevant for the model. We note that this penalty term may be viewed as a type of sparse group-lasso penalty [Simon et al., 2013]. The second term of the penalty function which is associated with the hyperparameter λ_2 is used to shrink towards zero the distances of very similar estimated cluster means. As a consequence, any two clusters with very similar cluster means will be forced to merge. This will reduce the initial assumed number of clusters G. Only clusters with very different means are expected to remain. We note that this penalty term may be viewed as a type of fused-lasso penalty [Tibshirani and Saunders, 2005].

2.3 The Bayesian lasso functional clustering model

Most of the literature concerning lasso-type penalization suggest using cross-validation in order to estimate and fix the value of the penalty parameters λ_1 and λ_2. Note that this procedure basically amounts to estimating the model for a given pair of optimal values of (λ_1, λ_2), ignoring the fact that the data (and the model) have been previously used to choose the pair (λ_1, λ_2). Another issue with cross-validation is its computational cost. This may be large for large datasets, and for complex models such as the one considered in this paper. Note that in order to find an optimal pair (λ_1, λ_2), a grid of values in the two-dimensional space of penalty parameters must be chosen. Therefore, a third issue with this procedure relates to how to choose the grid. If a simple uniform grid is to be chosen, then most of the time, the size of the grid would be too large. For example, for a 20 × 20 grid, one already needs to fit 400 models times the number of the cross-validation folds; if one performs a 5-fold cross-validation, the number of times one would need to fit the model would be 2000.

We suggest to choose the grid via techniques developed for the optimal design of experiments. One particular useful technique for computer experiments is Latin hypercube sampling (LHS). These are designs that try to fill in the search space much more efficiently than a uniform grid. For example, a uniform grid of size 10 × 10 may be too coarse to really find the optimal pair. But a LHS array of size 100 would cover the space of the penalty parameters in an efficient way. The LHS technique has been applied to many different computer models since 1975 (Steck et al. [1976], Iman et al. [1981a,b], Iman and Conover [1982a,b], Iman and Helton [1985], Wyss and Jorgensen [1998]).

Even though one could manage to reduce the number of model fits considerably by using the suggested LHS procedure, the cost of the search is sometimes still too high for large datasets. To alleviate this cost and to make sure we have obtained the optimal penalty parameters, we propose a model where the penalty function is part of the likelihood. This allows us to consider the pair (λ_1, λ_2) as model parameters, just as the rest of the parameters. Since the form of the penalty function is essential, we simply propose to normalize the penalized likelihood, that is, to make the penalty term a density. This solution requires finding the normalizing constant of the penalized likelihood function. The penalized likelihood function derived from the penalized log-likelihood in equation (5) can be expressed as $\mathbf{L}^{pen}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) = p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) \cdot e^{-P_\lambda(\vec{\boldsymbol{\mu}})}$. The complete expression of the density function $p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})$ and details on its computation are presented in Appendix A. We simply recall here that $p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})$ satisfies:

$$\begin{aligned} p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) &= p(\vec{\mathbf{Y}}, \vec{\mathbf{z}}, \vec{\gamma}^{\mathbf{z}}; \vec{\nu}, \vec{\boldsymbol{\mu}}, \vec{\mathbf{\Gamma}}, \mathbf{\Lambda}) \qquad (7) \\ &= \prod_{i=1}^{N} p(\mathbf{Y}_i, \mathbf{z}_i, \boldsymbol{\gamma}_i^{\mathbf{z}_i} \big| \nu_i, \vec{\boldsymbol{\mu}}, \vec{\mathbf{\Gamma}}, \mathbf{\Lambda}) \cdot p(\nu_i \big| \vec{\boldsymbol{\mu}}, \vec{\mathbf{\Gamma}}, \mathbf{\Lambda}) \cdot \prod_{g=1}^{G} p(\boldsymbol{\mu}_g) \cdot p(\mathbf{\Gamma}_g) \cdot p(\mathbf{\Lambda}) \\ &= \bar{p}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) \cdot \prod_{g=1}^{G} p(\boldsymbol{\mu}_g). \end{aligned}$$

where $\bar{p}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) = \prod_{i=1}^{N} p(\mathbf{Y}_i, \mathbf{z}_i, \boldsymbol{\gamma}_i^{\mathbf{z}_i} \big| \nu_i, \vec{\boldsymbol{\mu}}, \vec{\mathbf{\Gamma}}, \mathbf{\Lambda}) \cdot p(\nu_i \big| \vec{\boldsymbol{\mu}}, \vec{\mathbf{\Gamma}}, \mathbf{\Lambda}) \cdot \prod_{g=1}^{G} p(\mathbf{\Gamma}_g) \cdot p(\mathbf{\Lambda})$. The terms gathered in $\bar{p}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}))$ are known density functions of variables or parameters different from $\vec{\boldsymbol{\mu}}$. Therefore, in order to normalize $\mathbf{L}^{pen}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})$, we only need to normalize the penalty-induced prior on $\vec{\boldsymbol{\mu}}$, given by $\left[\prod_{g=1}^{G} p(\boldsymbol{\mu}_g) \cdot e^{-P_\lambda(\vec{\boldsymbol{\mu}})}\right]$. That is, we need to compute the integral

$$C(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu) = \left[\int \left(\prod_{g=1}^{G} \frac{e^{-\frac{1}{2}\boldsymbol{\mu}_g^{\intercal} \Gamma_\mu^{-1} \boldsymbol{\mu}_g}}{(2\pi)^{k/2} |\mathbf{\Gamma}_\mu|^{1/2}}\right) e^{-P_\lambda(\vec{\boldsymbol{\mu}})} d(\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_G)\right]. \qquad (8)$$

The normalized penalized log-likelihood is given by

$$\mathfrak{L}_c^{pen}(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi}) = \log[p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})] - P_\lambda(\vec{\boldsymbol{\mu}}) - \log C(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu). \qquad (9)$$

We refer to the model based on this normalized penalized log-likelihood as a *Bayesian lasso functional clustering model* or *Bayesian lasso FCM* for short. Note that the Bayesian lasso FCM is related but different to what has become known in the literature as Bayesian lasso [Park and Casella, 2008]. This latter procedure only penalizes individual parameters and does not involve group or fused-like penalization. A closer paradigm to the Bayesian lasso FCM would be the elastic-net penalty [Zou and Hastie, 2005] with sparse group and fused lasso penalties.

We use Monte Carlo numerical integration to estimate the integral in (8). We sample values $\{\vec{\boldsymbol{\mu}}_m\}_{m=1}^{M}$ according to a kG-Multivariate Normal distribution with mean zero and block-diagonal variance-covariance matrix with G blocks equal to $\mathbf{\Gamma}_\mu$. The sampling is done with respect to the prior distribution of $\vec{\boldsymbol{\mu}}$, which is exactly the term in parentheses in the integral above. Consequently, our estimator $\tilde{C}(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu)$ is given by

$$\tilde{C}(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu) = \left[\frac{1}{M} \sum_{m=1}^{M} e^{-\lambda_1 \sum_{j=1}^{k} (\sum_{g=1}^{G} |\mu_{g(m)}^{j}|) - \lambda_2 \sum_{g=1}^{G} \sum_{h=1}^{g} D^{ist}(\mu_{g(m)}, \mu_{h(m)})}\right]. \qquad (10)$$

In what follows we will refer to the unpenalized functional clustering model given by equations (2) and (3) as the unpenalized FCM. Also, we will refer to the penalized (unnormalized) functional clustering model given by equation (5) as the penalized FCM.

As mentioned in Section 2.1, the iterative EM algorithm is used to estimate the parameters of the model. The function $S(\mathbf{\Pi}|\bar{\mathbf{\Pi}})$ to be maximized using the EM algorithm for known values $\bar{\mathbf{\Pi}}$ of the parameters is defined by:

$$S(\mathbf{\Pi}|\bar{\mathbf{\Pi}}) = Q(\mathbf{\Pi}|\bar{\mathbf{\Pi}}) - P_\lambda(\vec{\boldsymbol{\mu}}) - \log C(\lambda_1, \lambda_2, \mathbf{\Gamma}_\mu), \tag{11}$$

where $Q(\mathbf{\Pi}|\bar{\mathbf{\Pi}}) = E_{\mathbf{W}|\mathbf{Y};\bar{\mathbf{\Pi}}}[\log p(\mathbf{Y}, \mathbf{W}; \mathbf{\Pi})]$. The observed data $\mathbf{Y}$ and the missing data $\mathbf{W}$ are as previously defined in Section 2.1. The computation of $Q(\mathbf{\Pi}|\bar{\mathbf{\Pi}})$ at the expectation step is identical to that of the unpenalized model given in Section 2.1. All the analytical developments leading to its calculus are the same as presented in Appendix A. The maximization step for all parameters in the model except for the cluster means $\boldsymbol{\mu}_g$ is identical to the one associated with the unpenalized model. The M-step equations for the cluster means $\boldsymbol{\mu}_g$ of the Bayesian lasso FCM are shown in Appendix B.

3 Model selection

In this section, we describe the steps toward the selection of the optimal model. Note that for the unpenalized FCM, model selection may be performed using criteria such as BIC or the Akaike information (AIC). However, these entail computing all possible models arising from all possible values of the two basic model parameters k and G. Instead, for the Bayesian lasso FCM, model selection consists only of the identification of the optimal number of functional principal components k^{opt} and the optimal number of clusters G^{opt} from the output of the EM algorithm. Note that for the penalized FCM, model selection is a two-stage procedure, since before the identification of the couple (k^{opt}, G^{opt}), one must identify the optimal values of the two tuning parameters $(\lambda_1^{opt}, \lambda_2^{opt})$ from a grid of proposed values. For the Bayesian lasso FCM, we have adopted the following steps for identifying the optimal values (k^{opt}, G^{opt}):

1. Postulate initial values for k and G: Let k^{pos} and G^{pos} be the postulated initial values of the number of functional components and the number of clusters, respectively. As mentioned in Section 2.2, the B-splines basis dimension q is not estimated but rather set manually. However, the value of q is not chosen at random. Some rules govern its choice. For a B-splines basis of order m_B (with polynomials of degree $d_B = m_B - 1$), the number of basis functions can be expressed as $q = m_B + i_B$ where i_B is the number of interior knots. And q must satisfy $q \geq m_B$. For example, q must be at least 4 in the case of cubic splines. Furthermore, the value of q must be large enough to ensure a significant number of interior knots that will be equally spaced within the range of the measurement time points $[\min t_{ij}, \max t_{ij}]$ $(1 \leq i \leq N; 1 \leq j \leq n_i)$ in order to span the individual curves. Also, as the columns of the (q,k)-dimensional matrix $\mathbf{\Theta}$ are orthonormal due to the orthogonality constraint $\mathbf{\Theta}^\mathsf{T}\mathbf{\Theta} = I_k$ [James and Sugar, 2003], the value of q must also satisfy $q \geq k^{pos}$. The values (q,k^{pos},G^{pos}) are used to initialize each single run of the EM-algorithm.

2. Perform the EM algorithm with ($\boldsymbol{\lambda_1}$,$\boldsymbol{\lambda_2}$): For each couple (λ_1, λ_2) and for each postulated (initial) values of the parameters (k^{pos}, G^{pos}), the Bayesian lasso FCM is fitted using the whole dataset. The parameters $\mathbf{\Theta}$, $\boldsymbol{\theta}_\mu$ and σ^2 are initialized by assuming a model with a single cluster. The cluster parameters such as $\boldsymbol{\mu}_g$, $\mathbf{\Gamma}_g$ and π_g as well as the cluster membership indicators $\mathbf{z}_i$ may be initialized by applying any clustering procedure to the scores $\boldsymbol{\alpha}_i$ yielded by the single-cluster model. In our experiments, we used the Gaussian model-based clustering procedure implemented in the *mclust package* (Fraley and Raftery [1999, 2003, 2006]).

3. Find ($\boldsymbol{\lambda_1^{opt}}, \boldsymbol{\lambda_2^{opt}}$): Step 2 is performed for each couple (λ_1, λ_2) in an appropriate grid (e.g., chosen via LHS). The optimal values $(\lambda_1^{opt}, \lambda_2^{opt})$ are defined as the couple (λ_1, λ_2) that maximizes the Bayesian lasso FCM log-likelihood. The next step is to determine k^{opt} and G^{opt} based on the estimators yielded by the Bayesian lasso FCM parameters associated with $(\lambda_1^{opt}, \lambda_2^{opt})$.

4. Find (k^{opt},G^{opt}): The optimal values of G and k are identified by examining, respectively, the matrix of the between distances of the cluster means, $\boldsymbol{D_M}$, and the vector of elements $v_j = \left[\sum_{g=1}^{G} \left|\mu_g^j\right|\right]$ with $j = 1, ..., k^{pos}$, which will be denoted by $\boldsymbol{V_M} = (v_j)_{j=1}^{k^{pos}}$.

- **Determining** k^{opt}: The optimal number of functional principal components is obtained by reducing k^{pos}. The elements associated with very small values of v_j are dropped from the model. For that purpose, two criteria are proposed. The optimal k is set to minimum of the values provided by these criteria. The first criterion is inspired by the notion of inertia in classical principal components analysis. Recall that the inertia of a factor corresponds to the information it carried. In our setup, the inertia of a the j^{th} component is associated with its v_j value. Similarly to the criterion of cumulative proportion of total inertia, we search among the top ranked v_j values for the minimum number of principal components contributing to at least 80% of the cumulative sum; that is, we look for the smallest $k^* \leq k^{pos}$ such that $\sum_{j=1}^{k^*} v_{(j)} \geq 0.80 \sum_{j=1}^{k^{pos}} v_j$, where $v_{(1)} \geq v_{(2)} \geq .... \geq v_{(k^{pos})}$ are the ranked statistics associated with the components of the vector $\boldsymbol{V_M}$. The other criterion is based on an approximate multiple testing procedure. Consider $\{\mu_1^j, \ldots, \mu_G^j\}$, as a sample, and $v_j = \sum_{g=1}^{G} \left|\mu_g^j\right|$, as an associated statistic. We would like to test the null hypothesis of zero posterior expectation $E(\mu_g^j) = 0$, for all $g = 1, \ldots, G$. As a heuristic, we suppose that under the null hypothesis the posterior of each μ_g^j follows a mean-zero Normal distribution with a common variance σ_k^2. Under the null hypothesis, v_j is distributed as a sum of G independent half-Normal$(0, \sigma_k^2)$. The half-normal distribution is a fold at the mean of an ordinary normal distribution with mean zero. Although the density associated with v_j is not known in closed form, we can easily estimate percentiles from its distribution by Monte Carlo simulation when σ_k^2 is known.

Under the null hypothesis, $E(|\mu_g^j|) = \sigma_k\sqrt{2/\pi}$. Therefore, up to a constant, the mean of the components of the vector $\boldsymbol{V_M}$ is an estimate of σ_k. In practice, we expect only a few components to be negligible; so the estimate of σ_k may be taken from only the smallest v_js, or even just the minimum of the v_js. Since k^{pos} simultaneous tests need to be performed (one for each v_j), we apply a Bonferroni correction and work with a threshold $R_{\alpha/k}$ so that we do not reject the null hypothesis if the observed value of v_j is smaller or equal to this threshold. Note that the threshold is given by the equation $P(T \leq R_{\alpha/k}) = \alpha/k$, where T is distributed as σ_k times the sum of G independent half-Normal$(0,1)$. That is, $R_{\alpha/k} = \sigma_k q_{\tilde{T},\alpha/k}$, where $q_{\tilde{T},\alpha}$ is the $100\,\alpha^{th}$ percentile of the sampled distribution of the sum of G independent half-Normal$(0,1)$.

• **Determining** G^{opt}: We recall that the fundamental idea in the identification of the optimal number of clusters is the merging of clusters with identical characteristics, that is, with very small between cluster mean distances . We suppose that the distances between vectors given by the matrix $\boldsymbol{D_M} = (D_{gh})_{1\leq g,h\leq G}$ are Euclidean distances. Our heuristic assumes that under the null hypothesis of zero distance (that is $\boldsymbol{\mu}_g = \boldsymbol{\mu}_h$), the posterior distribution of each pair of means is given by $(\boldsymbol{\mu}_g - \boldsymbol{\mu}_h) \sim \mathcal{N}_k(0, \sigma_G^2 I_k)$ with a common scale parameter σ_G^2. In this case, we have $D_{gh}^2 = \sum_{l=1}^k (\mu_g^l - \mu_h^l)^2 \sim \sigma_G^2 \chi_k^2$. We estimate the scale parameter σ_G^2 as the mean of the squared distances $\hat{\sigma}_G^2 = \frac{2}{G(G-1)} \sum_{g<h}^k D_{gh}^2$, and plug this estimator in the above equation. As in the case of k^{opt}, in practice, we only use the smallest elements D_{gh} in the estimation of $\hat{\sigma}_G^2$. Using the Bonferroni correction for multiple testing, the null hypothesis is not rejected if the observed value D_{gh}^2 is not larger than $\hat{\sigma}_G^2$ times the lower $100(\alpha/G)^{th}$ percentile of a χ_k^2 distribution. Note that for simplicity, we have assumed mutual independence between the distances.

4 Simulation study

We conduct a simulation study to examine the performance of the proposed methodology. We investigate specifically the ability of the method to reproduce and cluster original curves by correctly estimating the key parameters from postulated values. Following the described model selection procedure, the simulation study also concentrates on identifying the most relevant threshold to consider for the determination of G^{opt}.

Simulations setup

Various curves are generated based on different values of the sample size $N \in \{100, 500, 900, 4000\}$, the spline basis dimension $q \in \{10, 12, 14, 15, 20\}$, the number of functional principal components $k \in \{2, 4, 5, 6, 8\}$ and the number of clusters $G \in \{3, 6, 9, 15, 20, 40\}$. Overall, for each combination of (N, q, k, G), the parameters of the model are generated at random according to the prior distributions assumed in the proposed model. We consider individual curves measured at 12 time periods. This choice reflects the size of data from the microarray transcriptome of rats exposed to cigarette smoke (see Section 6). For the simulations, 50 couples (λ_1, λ_2) were randomly obtained using a Latin hypercube sampling scheme with values in the rectangle $[0, 50] \times [0, 50]$. As the B-splines basis dimension q is not estimated, its value is set to the one used to generate the curves. We postulate $k^{pos} = 8$ for the number of functional principal components. For the sake of efficiency in the simulations, the postulated values for the number of clusters G^{pos} depended on the true number of clusters G. The corresponding set of values are given in Table 1.

Table 1: Values of the postulated number of clusters $G^{(pos)}$ according to the true number of clusters G

G	G^{pos}					G	G^{pos}				
3	12	10	8	6	4	15	26	24	22	20	18
6	16	14	12	10	8	20	30	28	26	24	22
9	19	17	15	13	11	40	50	48	46	44	42

Simulation analysis tools

The simulation is based on 12 datasets generated as described above, each one corresponding to a specific (and different) combination of (N, q, k, G). For each dataset, the couples (λ_1, λ_2) were generated and the procedure was launched for each value of G^{pos}, with $k^{pos} = 8$. The quality of the results is assessed by comparing the partitions (clustering) created by the model and the original (true) cluster memberships. The comparison was made through the Adjusted Rand Index (ARI) (Rand [1971], Hubert and Arabie [1985]). A perfect agreement between the two partitions yields an ARI score of 1. The closest the score is to 1, the more similar the partitions are. The ARI has become the standard measure of comparison in the statistical literature on clustering. Three different thresholds were proposed to estimate the optimal number of clusters. As mentioned in the previous section, we consider only the smallest distances for the estimation of $\hat{\sigma}^2_G$. The first criterion, $\mathbf{1}\boldsymbol{low}$, uses the smallest distance in $\boldsymbol{D_M}$; the second criterion, $\mathbf{25}\boldsymbol{low}$, uses the mean of the distances falling below the first quartile of the distances in $\boldsymbol{D_M}$; and the third criterion, $\mathbf{50}\boldsymbol{low}$, uses the mean of the distances falling below the median of the distances in $\boldsymbol{D_M}$. Therefore, in addition to the model parameters estimates, the relevant quantities yielded by each simulation run are $\{k^{opt}, G^{opt}_{1low}, \text{ARI}_{1low}, G^{opt}_{25low}, \text{ARI}_{25low}, G^{opt}_{50low}, \text{ARI}_{50low}\}$, where the subindex represents the criterion used to choose G^{opt}. In order to calibrate the ARI index with the difficulty of the problem, we also report a measure of data complexity as presented by Chen et al. [2002]. Let N be the total number of curves in the data, n_g be the number of curves in cluster g, and MC_g be the mean curve of cluster g. Consider the following measures of *Homegeneity* and

Separation given respectively by

$$H = \frac{1}{N}\sum_{i=1}^{N} D^{ist}(Y_i, MC_{z_i}), \qquad S = \frac{1}{\sum_{g\neq h}^{G} n_g n_h}\sum_{g\neq h}^{G} n_g n_h D^{ist}(MC_g, MC_h).$$

The homogeneity is calculated as the average distance between each curve and the mean curve of the cluster it belongs to. It reflects the compactness of the clusters. The separation is calculated as the weighted average distance between the cluster mean curves. It reflects the overall distance between clusters. As the indices H and S are closely related to respectively within-cluster and between-cluster variances, the similarity ratio $Ratio = 1 - \left(\frac{N}{N-1}\right)\left(\frac{H}{H+S}\right)$, serves as a measure of homogeneity: datasets with large similarity ratios are easier to cluster than those with small similarity ratios.

Simulation results

The first element of the simulation study is the comparison of the three threshold criteria proposed to estimate the optimal number of clusters. For each dataset, we computed, for each threshold criterion, the average of the ARIs from the five different postulated G^{pos}. The results of an analysis of variance indicate that the criteria are significantly different. As shown in Figure 2, the criterion ***1low*** appears to perform best. In Figure 3, we compare the similarity ratios and the criterion ***1low*** average ARI_{1low} for every dataset. In addition, in this figure, we also compare the performance of the current model with the unpenalized FCM. This

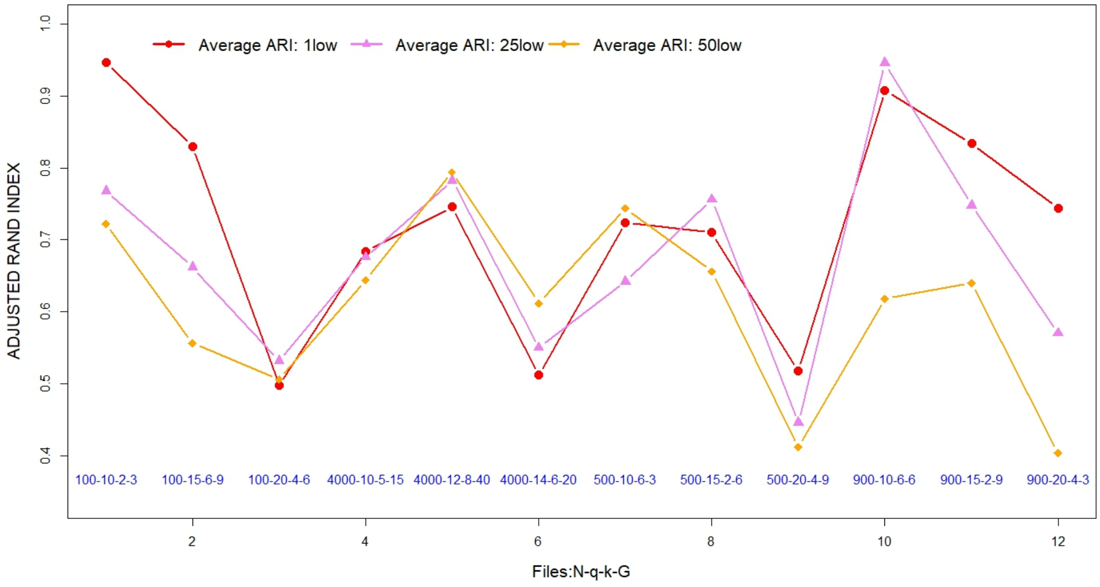


Figure 2: Comparison of the three criteria for model selection.

latter model uses the Bayesian information criterion (BIC) to choose the model parameters k, q and G. Because this model evaluates all possible models in a grid of values of (k, q, G), its performance might be better than that of the Bayesian lasso FMC. However, for the same reason, its computational cost is much larger. The boxplots in Figure 3 are associated with the ARI values obtained from the five different postulated G^{pos}. The figure shows that the results from the two different FCMs are comparable. However, the Bayesian lasso FCM has found the clusterings with much less computational cost. Also, note that the trend in the similarity ratio is also depicted in the ARI averages. This observation indicates that the clustering performance of the models are highly related to the degree of complexity of the data.

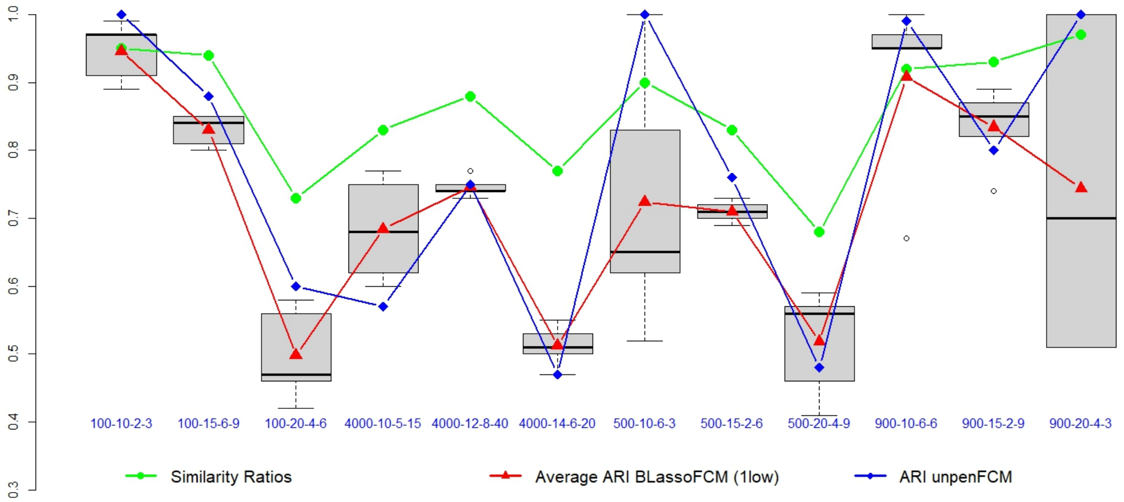


Figure 3: Similarity ratio and model performance.

Another element analyzed in the simulation is the impact of the proposed number of cluster G^{pos}. The question addressed here is whether there is a significant difference in the clustering results if G^{pos} is far from or close to the true number of clusters. The answer would give an indication on how to select G^{pos} in practice. For that purpose, we draw in Figure 4 a scatterplot of the values $(\sqrt{G^{pos}} - \sqrt{G})$ (representing the gap between proposed and true G) against the corresponding ARI_{1low} criterion. There is no structure nor trend observable from this figure. The conclusion is that there appears to be no relationship between the clustering performance and the proposed G in the algorithm. No matter how close or far G^{pos} is to the true number of clusters, the model performs similarly.

Finally, we evaluate the capacity of the model to replicate the real number of clusters in the

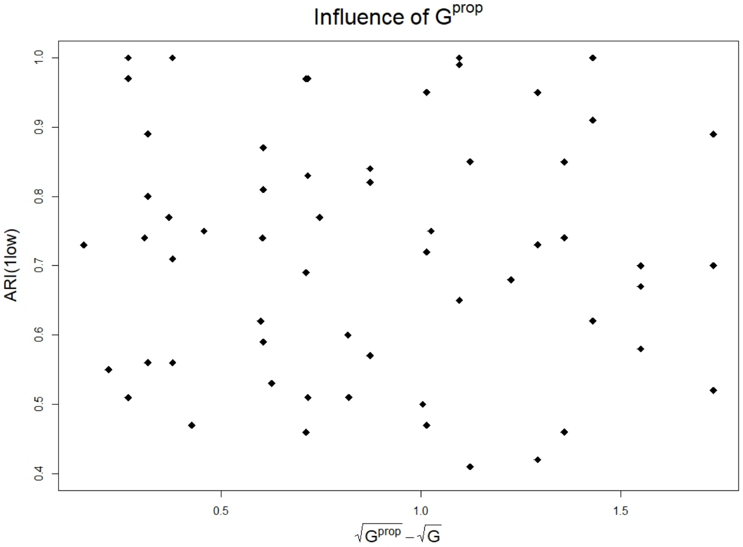


Figure 4: Influence of postulated G

data. Consider the gap between the estimated and true G given by $(\sqrt{G^{est}} - \sqrt{G})$. For the unpenalized FCM, the gap is a single value for each dataset. For the Bayesian lasso FCM, the gap variable is represented by the set of G^{est} obtained through the postulated values for G. This is reflected by the boxplots in Figure 5. Note that the average gap values for the Bayesian lasso FCM are very small and comparable to the ones from the unpenalized functional clustering model.

5 Comparison study

We compared the performance of the Bayesian lasso FCM with five other functional clustering models: `fitfclust` of James and Sugar [2003], `iterSubspace` of Chiou and Li [2007], `distFPCA` of Peng and Müller [2008], `funHDDC` of Bouveyron and Jacques [2011], `fscm` of Giacofci et al. [2013], and `funclust` of Jacques and Preda [2013]. We use the implementation of these models given in the R package `funcy` [Yassouridis, 2017]. Unlike Bayesian lasso FCM, the models in `funcy` do not estimate automatically the number of clusters. So to make the comparison, we run the methods in `funcy` with several proposed G values and kept the value that maximized BIC. The results are shown in Figure 6. In the figure, the results associated with the unpenalized FCM method are denoted by `unpenFCM`, and the results associated with the Bayesian lasso FCM are denoted by `BLassoFCM`. It is clear from this figure that the unpenalized FCM and the Bayesian lasso FCM are the best performers.

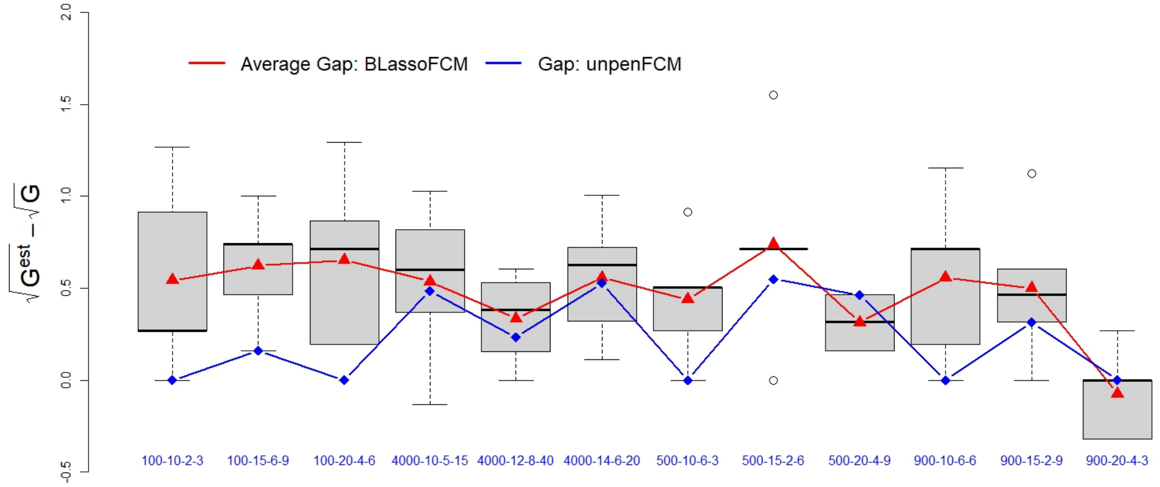


Figure 5: Comparison of True and estimated number of clusters

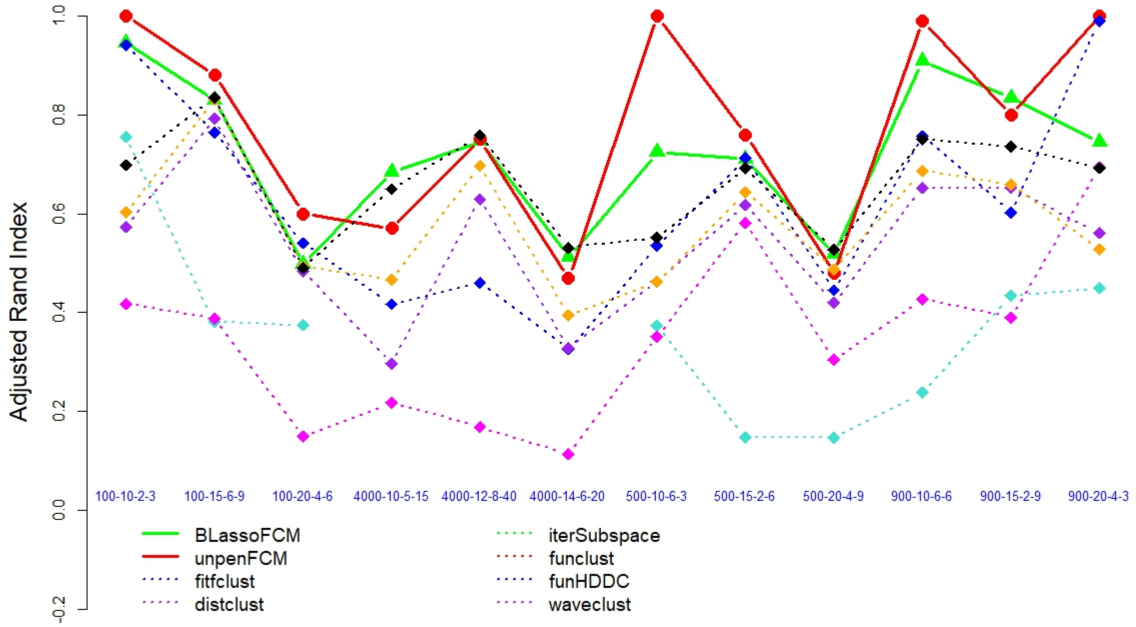


Figure 6: Comparison of functional clustering methods

Running example.
We have also obtained similar results with real dataset experiments (not shown here). As an illustration, we come back to the yeast cycle data introduced earlier in Section 2.1 Cho et al. [1998]. Recall that experts estimate that the data reveal fives phases. However, to our knowledge, there is no clustering procedure that can automatically reproduce these phases adequately. Our model suggest a four-cluster partition with an associated ARI of slightly over 0.47. These results are highly comparable to those obtained by other studies on the same data set. Indeed, The Potts model clustering of Murua et al. [2008] yielded nine clusters with an ARI of 0.45. Yeung et al. [2001] analyzed the same subset of these data using model-based clustering based on Gaussian mixtures developed in Banfield and Raftery [1993]. They reported four clusters with an ARI of about 0.43. The bottom two rows in Figure 7 show the four mean curves associated with the four-cluster solution. The figure displays the observed and the estimated mean curves. For comparison purposes, we also show in Figure 7 (see top three rows) the five mean curves associated with the five original clusters proposed by Cho et al. [1998].

Observe that the mean curves associated with the four estimated clusters are very similar to the first four clusters found by Cho et al. [1998]. The fifth cluster of Cho et al. [1998] lies between clusters two and four of the estimated clusters. The covariance structure of the four estimated clusters is displayed in Figure 1. Observe that there is high correlation between time points close to valleys and peaks in the mean curves. One can also observe a slightly negative correlation for points further away from the valleys and peaks. The estimation of the error-term degrees of freedom is $\nu = 3$. The middle right panel of Figure 7 displays the distribution of the ν_i variables associated with each observation. Recall that the error terms ϵ_i were modeled as a convolution of Normal and Inverse-χ^2 distributions, so that small values of ν_i give evidence of non-nomally distributed errors.

6 Chronic obstructive pulmonary disease

We applied the Bayesian lasso functional clustering model to shed light into the initial molecular events linked to chronic obstructive pulmonary disease (COPD). The dataset, described previously in the introduction, relates to time-course genetic expression difference between tobacco-smoke exposed rats (the treatment group) and a control group of non-exposed rats (Stevenson et al. [2007]). The dataset comes from the project GEO GSE7079 (Gene Expression Omnibus [2007]) and is related to a study of molecular changes due to the exposure of rats to tobacco smoke. It is a time-course data with 12 day time-points: 1, 3, 5, 14, 21, 28, 42, 56, 84, 112, 182 and 238. Probesets (genes) without any GO annotation were discarded (Gene Ontology Consortium [1999-2015]). The 3464 probesets considered in our study corre-

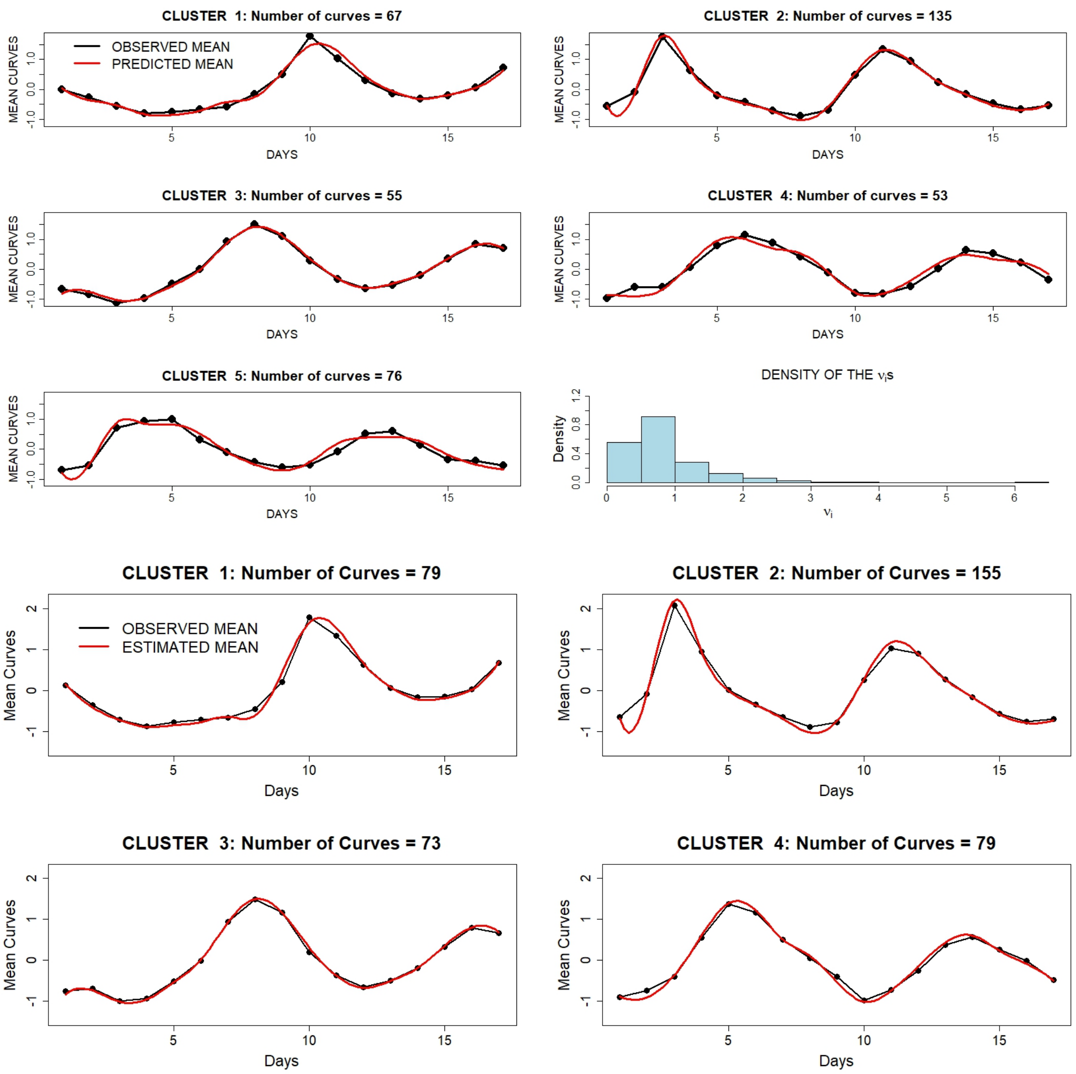


Figure 7: Yeast cycle data. Top three rows: observed and model-estimated mean curves for the five clusters found by Cho et al. [1998]. Right panel in third row: distribution of the ν_i associated with the error term of the FCM model. Bottom three rows: Observed and model-estimated mean curves for the four clusters yielded by FCM

spond to 39.4% of the original 8799 probesets in the dataset.

Analysis

We set the initial proposed number of clusters G^{pos} to $50, 20$, and 10 clusters. These choices led respectively to models with 22, 11 and 7 clusters. Despite the difference in the number of clusters, all three partitions are very similar in the sense that those partitions with smaller number of clusters are basically formed by merging of clusters in the larger partitions. Figure 8 displays the cluster means from all three partitions. The bottom row shows the estimated cluster-specific mean curves, while the plots on the top row show the two-dimensional graphical representation of the functional principal components scores. Recall that the cluster mean curves are given by linear combinations of the functional principal components. For these particular data, these two-dimensional representations are exact because the estimated dimension k of the curves is exactly 2.

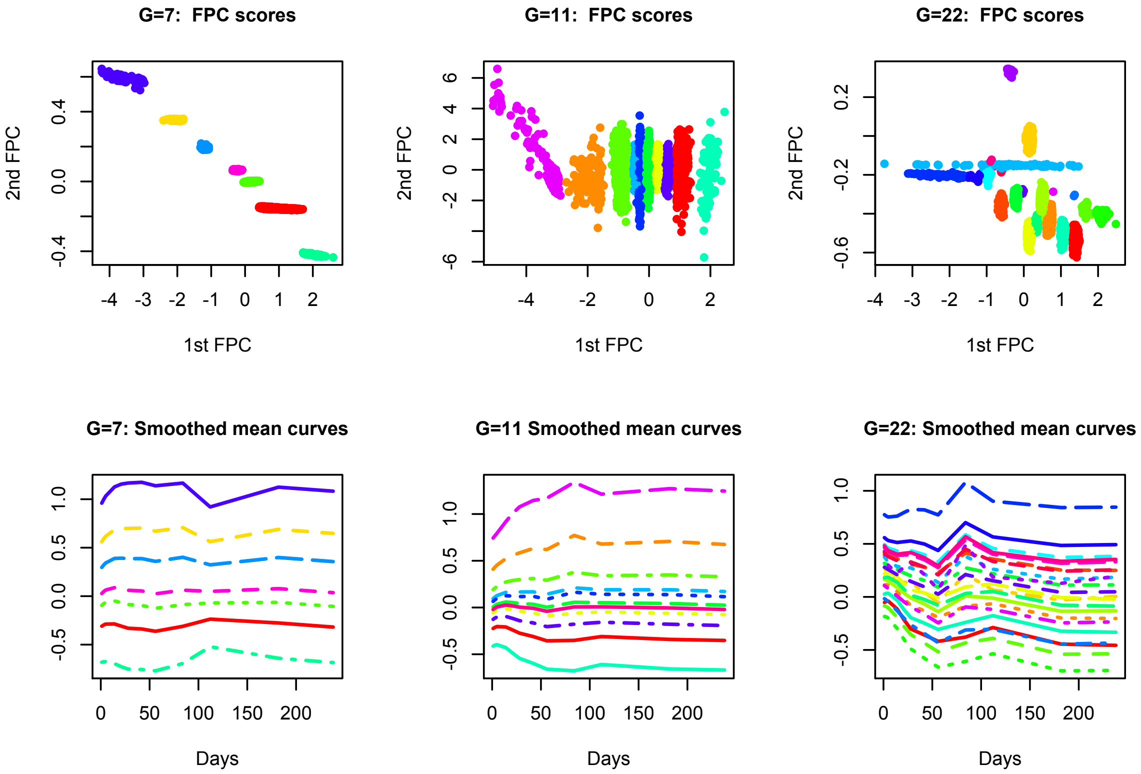


Figure 8: The exposure to tobacco smoke data. Cluster mean curves for the three partitions found by the Bayesian lasso functional clustering model. The top row displays the clusters in the two-dimensional space of functional principal components (FPC) scores. The bottom panel shows the cluster mean-curves.

During statistical analysis, all genes tested for their abundance levels of corresponding mRNAs (expression) are typically considered as independent entities. While most resultant proteins (the role of genes is to produce proteins) are independent molecules, they typically act as to form chains of reactions (e.g., membrane receptors and subsequent intracellular signaling), or multi-protein complexes. In such cases, all individual elements are required to be present to allow proper biological functioning. In many cases, it has been observed that the expression in such subsets of proteins is regulated in a similar way. Also, gene clusters differentially regulated at specific biological conditions frequently contain pathway footprints that lead to enrichments in gene-ontologies (such as the hierarchical GO ontology). For this reason, testing enrichment of GO-ontologies and gene-pathway analysis are considered complementary approaches to statistical testing of differential expression. Gene-pathways are typically built based on information in databases of protein-protein interactions. Large-scale screening experiments in given organisms, like yeast, allow researchers to determine which proteins form compact aggregate structures while exercising a given biological role. Such databases are typically combined with outcome from automated mining of biochemical publications.

In our case, the resulting gene clusters were inspected and characterized for enrichment with the GO-ontologies (gene functions). Clusters were judged interesting according to their patterns of expression profile, and for their high content of GO-terms associated with smoke-exposure. These clusters were then submitted to protein-protein network analysis in order to look for additional enrichment of expression patterns, or functionalities, across selected elements of the networks obtained. More specifically, the three partitions of genes found by the Bayesian lasso functional clustering model were inspected for enrichment of functionalities with the DAVID platform (Huang et al. [2009], Maere et al. [2005], National Institute of Allergy and Infectious Diseases, NIH [2017]). This partition revealed several clusters highly enriched in functions previously attributed to acute and chronic exposure to cigarette smoke: immune response/immune system (clusters 1, 8, and partially 12), inflammation (clusters 15, 16), and apoptosis (clusters 3, 5, 6, 11). Surprisingly, these latter clusters, which are associated to late/prolonged exposure with cigarette smoke, do not share general expression patterns such as global up or down-regulation profiles. However, clusters 15 and 16, associated with early exposure, do share a common upregulated expression pattern. In the 7-cluster partition, four clusters may be characterized by enrichment of gene functions directly related to early and late phases of tobacco smoke exposure. Among them, there is a cluster which despite its large size still has an interesting expression profile and excellent GO ontology enrichment scores: it represents genes that are gradually and increasingly repressed during the entire process of exposure to smoke. Curiously, all clusters enriched in gene functions associated with early phase of smoke exposure are also enriched in functions associated to long/chronic exposure. Among these, genes in cluster 2 do not show major expression changes during the late phase. Probably, these genes are not genes triggering chronic symptoms, but are genes that when activated « set the stage », that is, they may be associated with acute sensitivity for developing symptoms at a prolonged smoke exposure. In contrast, the cluster of genes specifically upregulated in the late phase has a more fuzzy profile, that is, there is no simple or clear tendency in the gene expressions. The 11-cluster partition presents three clusters identified as predominantly characterizing gene-functions associated with exposure to smoke. In all three clusters, the expression profiles are in perfect accord to whether genes functions are associated with early or late phases of smoking.

7 Conclusion

In this paper, we introduced a model-based Bayesian lasso functional clustering method for the analysis of longitudinal data. The model combines dimension reduction and clustering through functional principal component analysis, and model-based clustering. Model selection is done through a lasso driven prior for the cluster means. Latin Hypercube Sampling was used to efficiently explore the space of penalty parameters.

The analysis of gene expression from smoke exposure showed that many deregulation events are associated with relevant gene-functions. This suggests that gene-repression may be a very common effect associated with biological effects of smoke exposure. We note that gene-repression is typically more difficult to find by classical data analysis approaches, and in consequence, it is frequently less regarded. The case of upregulated genes may thus be more punctual for specific aspects. In summary, one may conclude that the clustering approach allowed for identification of large groups of gradually deregulated genes that otherwise might be difficult to capture using traditional statistical approaches such as multiple testing of two groups (e.g., smoke-exposed versus control groups). We recall that we have introduced a different approach in the analysis of these data by considering all time points without introducing bias, that is, without restricting our study to the most extreme changes like in the initial study [Stevenson et al., 2007]. This unbiased view lets us understand that the biology of stress response to smoke exposure may be more complex that previously thought, specially when compared to the simplistic view of the initial study. Furthermore, this novel interpretation is in agreement with other integrative studies, emphasizing that during a cell's lifetime many genes are used in many different situations, such as in both acute and chronic phases of stress resulting from tobacco exposure.

In conclusion, unbiased analysis methods (such as our Bayesian lasso FCM) allow a more general view of temporal or longitudinal processes that may be easily overseen by traditional approaches which focus on isolating the most extreme points first. Of course, such "holistic" approaches favour an integrated view of underlying processes and the complexity of regulatory systems that have evolved over long periods of evolution in biology.

References

F Adjogou. *Analyse statistique de données fonctionnelles à structures complexes.* PhD thesis, Université de Montréal, 2017.

J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

C. Bouveyron and C. Brunet. Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, 71:52–78, 2014.

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

G. Chen, S. A. Jaradat, N. Banerjee, T. S. Tanaka, M. S. H. Ko, and M. Q. Zhang. Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. *Statistica Sinica*, pages 241–262, 2002.

J.-M. Chiou and P.-L. Li. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society. Series B*, 69(4):679–699, 2007.

R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cel L*, 2:65–73, 1998.

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

C. Fraley and A. E. Raftery. Mclust : Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306, 1999.

C. Fraley and A. E. Raftery. Enhanced software for model-based clustering, density estimation, and discriminant analysis : Mclust. *Journal of Classification*, 20:263–286, 2003.

C. Fraley and A. E. Raftery. Mclust version 3 for r : Normal mixture modeling and model-based clustering. Technical Report 504, 2006. Department of Statistics, University of Washington.

Gene Expression Omnibus. Chronic rat exposure to cigarette smoke. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7079`, 2007.

Gene Ontology Consortium. GO. `http://www.geneontology.org/`, 1999-2015.

M. Giacofci, S. Lambert-Lacroix, G. Marot, and F. Picard. Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics*, 69(1):31–40, 2013.

G. H. Golub and C. F. Van Loan. *Parameter estimation.* Johns Hopkins, Baltimore, MD, 1996.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protoc.*, 4(1):44–57, 2009.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

R. L. Iman and J. C. Helton. A comparison of uncertainty and sensitivityanalysis techniques for computer models. NUREGKR-3904, SAND84-1461., 1985. Albuquerque, NM: Sandia National Laboratories.

R.L. Iman and W.J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics*, B11(3):311–334, 1982a.

R.L. Iman and W.J. Conover. Sensitivity analysis techniques: Self-teaching curriculum. Nuclear Regulatory Commission Report, NUREG/CR-2350, Technical Report SAND81-1978, Sandia National Laboratories, Albuquerque, NM., 1982b.

R.L. Iman, J.C. Helton, and J.E. Campbell. An approach to sensitivity analysis of computer models, part 1. introduction, input variable selection and preliminary variable assessment. *Journal of Quality Technology*, 13(3):174–183, 1981a.

R.L. Iman, J.C. Helton, and J.E. Campbell. An approach to sensitivity analysis of computer models, part 2. ranking of input variables, response surface validation, distribution effect and technique synopsis. *Journal of Quality Technology*, 13(4):232–240, 1981b.

J. Jacques and C. Preda. Funclust: A curves clustering method using functional random variable density approximation. *Neurocomputing*, 112:164–171, 2013.

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

G. James and C. A. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98:397–408, 2003.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38: 963–974, 1982.

S. Maere, K. Heymans, and M. Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005.

A. Murua, L. Stanberry, and W. Stuetzle. On potts model clustering, kernel k-means and density estimation. *Journal of Computational and Graphical Statistics*, 17:629–658, 2008.

National Institute of Allergy and Infectious Diseases, NIH. DAVID bioinformatics resources. `https://david.ncifcrf.gov/`, 2017.

W. Pan and X. Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8:1145–1164, 2007.

T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103:681–686, 2008.

J. Peng and H. G. Müller. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, 2(3):11056–1077, 2008.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

A. Same, F. Chamroukhi, G. Govaert, and P. Aknin. Model-based clustering and segmentation of times series with changes in regime. *Advances in Data Analysis and Classification*, 5(4): 301–322, 2011.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

G. P. Steck, R. L. Iman, and D. A. Dahlgren. Probablistic analysis of loca , annual report for 1976. SAND76-0535, Sandia National Laboratories, Albuquerque, NM., 1976.

C.S. Stevenson, Docx C., R. Webster, C. Battram, D. Hynx, J. Giddings, P.R. Cooper, P. Chakravarty, I. Rahman, J.A. Marwick, P.A. Kirkham, C. Charman, D.L. Richardson, N.R. Nirmala, P. Whittaker, and K. Butler. Comprehensive gene expression profiling of rat lung reveals distinct acute and chronic responses to cigarette smoke inhalation. *Am. J. Physiol Lung Cell Mol. Physiol.*, 293(5):L1183–93, 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society Series B*, 73:273–282, 2011.

R. Tibshirani and M. Saunders. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005.

S. Ullah and C. F. Finch. Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 1471-2288:13–43, 2013.

G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer series in statistics. New York, 2000.

S. Wang and J. Zhou. Variable selection for model-based high dimensional clustering and its application to microarray data. *Biometrics*, 64:440–448, 2008.

G. D. Wyss and K. H. Jorgensen. A user's guide to lhs: Sandia's latin hypercube sampling software. AND98-0210 Distribution Unlimited Release Category UC-505, 1998. Risk Assessment and Systems Modeling Department Sandia National Laboratories.

Christina Yassouridis. *funcy: Functional Clustering Algorithms*, 2017. URL `https://CRAN.R-project.org/package=funcy`. R package version 0.8.6.

K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987, 2001.

X. Zhao, J.S. Marron, and M.T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, 14:789–808, 2004.

H. Zou and T. Hastie. Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.

A The likelihood and EM updating equations for the unpenalized model

The log-likelihood $\log[p(\mathbf{Y},\mathbf{W};\boldsymbol{\Pi})]$ is

$$
\begin{aligned}
&\sum_{i=1}^{N}\left\{\begin{array}{l}
-\frac{n_i}{2}\log(\nu_i\sigma^2)-\frac{1}{2\nu_i\sigma^2}\left\|\mathbf{Y}_i-(\mathbf{B}_i\boldsymbol{\theta}_\mu+\mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{z}_i}+\mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\gamma}_i^{\mathbf{z}_i})\right\|^2\\
-\frac{1}{2}\log(\left|\boldsymbol{\Gamma}_{\mathbf{z}_i}\right|)-\frac{1}{2}\boldsymbol{\gamma}_{i,\mathbf{z}_i}^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1}\gamma_i^{\mathbf{z}_i}+\sum_{g=1}^{G}Z_{ig}\log(\pi_g)\\
+\frac{\nu_o}{2}\log(\frac{\nu_o}{2})-\log[\Gamma(\frac{\nu_o}{2})]-(1+\frac{\nu_o}{2})\log(\nu_i)-\frac{\nu_o}{2\nu_i}
\end{array}\right\}\\
&+\sum_{g=1}^{G}\left\{\begin{array}{l}
-\frac{1}{2}\log(\left|\boldsymbol{\Gamma}_\mu\right|)-\frac{1}{2}\boldsymbol{\mu}_g^T\boldsymbol{\Gamma}_\mu^{-1}\boldsymbol{\mu}_g+\frac{m}{2}\log(\left|(m-k-1)\mathbf{D}\right|)\\
-\frac{(m+k+1)}{2}\log(\left|\boldsymbol{\Gamma}_g\right|)-\frac{(m-k-1)}{2}\operatorname{trace}[\mathbf{D}\boldsymbol{\Gamma}_g^{-1}]
\end{array}\right\}\\
&+\left\{\ \frac{km}{2}\log(m-k-1)-\frac{(m+k+1)}{2}\log(\left|\boldsymbol{\Gamma}_\mu\right|)-\frac{(m-k-1)}{2}\operatorname{trace}[\boldsymbol{\Gamma}_\mu^{-1}]\ \right\}\\
&+\sum_{j=1}^{k}\left\{\ +\frac{m}{2}\log(\frac{m}{2})-\log[\Gamma(\frac{m}{2})]-(1+\frac{m}{2})\log(d_{jj})-\frac{m}{2d_{jj}}\ \right\}\\
&+\left\{\ \alpha_\sigma\log(\beta_\sigma)-\log[\Gamma(\alpha_\sigma)]-(\alpha_\sigma+1)\log(\sigma^2)-\frac{\beta_\sigma}{\sigma^2}\ \right\}\\
&+\left\{\ -\log[B(a_1,...,a_G)]+\sum_{g=1}^{G}(a_g-1)\log(\pi_g)\ \right\}\\
&+\mathcal{C}
\end{aligned}
$$

where $\mathcal{C}$ is the normalizing constant, and $B(a_1,...,a_G)=B(\boldsymbol{a})$ is the multivariate Beta function which can be expressed in terms of the Gamma function $\Gamma(\cdot)$ as $B(\boldsymbol{a})=\frac{\prod_{g=1}^{G}\Gamma(a_g)}{\Gamma(\sum_{g=1}^{G}a_g)}$.

We can rewrite the last expression of $\log[p(\mathbf{Y},\mathbf{W};\boldsymbol{\Pi})]$ as $\log[p(\mathbf{Y},\mathbf{W};\boldsymbol{\Pi})]=\mathcal{L}+\mathcal{H}$ where $\mathcal{L}$ groups the terms depending on the individuals $i=1,\ldots,N$:

$$
\begin{aligned}
\mathcal{L}=\sum_{i=1}^{N}l_i(\vec{\boldsymbol{\mu}},\vec{\boldsymbol{\Gamma}},\boldsymbol{\Lambda})=\sum_{i=1}^{N}&-\frac{n_i}{2}\log(\nu_i\sigma^2)-\frac{1}{2\nu_i\sigma^2}\left\|\mathbf{Y}_i-(\mathbf{B}_i\boldsymbol{\theta}_\mu+\mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{z}_i}+\mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\gamma}_i^{\mathbf{z}_i})\right\|^2\\
&-\frac{1}{2}\log(\left|\boldsymbol{\Gamma}_{\mathbf{z}_i}\right|)-\frac{1}{2}\boldsymbol{\gamma}_{i,\mathbf{z}_i}^{\mathsf{T}}\boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1}\boldsymbol{\gamma}_i^{\mathbf{z}_i}+\sum_{g=1}^{G}z_{ig}\log(\pi_g)\\
&+\frac{\nu_o}{2}\log(\frac{\nu_o}{2})-\log[\Gamma(\frac{\nu_o}{2})]-(1+\frac{\nu_o}{2})\log(\nu_i)-\frac{\nu_o}{2\nu_i}
\end{aligned}
$$

and $\mathcal{H}$ groups the remainder terms.

At iteration $(t+1)$ of the EM algorithm, for $\mathbf{\Pi}^{(t)}$ fixed, the function Q to be maximized is

$$\begin{aligned}
Q(\mathbf{\Pi}|\mathbf{\Pi}^{(t)}) &= E_{\mathbf{W}|\mathbf{Y};\mathbf{\Pi}^{(t)}}\left[\log p(\mathbf{Y},\mathbf{W};\mathbf{\Pi})\right] = E_{\mathbf{Z},\vec{\boldsymbol{\gamma}}^{\mathbf{z}}|\mathbf{Y},\vec{\nu}^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)}}\left[\mathcal{L}+\mathcal{H}\right] \\
&= \sum_{i=1}^{N} E_{\mathbf{z}_i,\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{Y}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)}}\left[l_i(\vec{\boldsymbol{\mu}},\vec{\mathbf{\Gamma}},\mathbf{\Lambda})\right] + \mathcal{H} \\
&= \sum_{i=1}^{N} E_{\mathbf{z}_i|\mathbf{Y}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)}}\left\{E_{\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{z}_i,\mathbf{Y}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)}}[l_i(\vec{\mu},\vec{\Gamma},\mathbf{\Lambda})]\right\} + \mathcal{H}. \qquad \text{(A.1)}
\end{aligned}$$

Let $m_i(\mathbf{z}_i,\vec{\mu},\vec{\Gamma},\mathbf{\Lambda}) = E_{\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{z}_i,\mathbf{Y}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)}}[l_i(\vec{\mu},\vec{\Gamma},\mathbf{\Lambda})]$. In order to compute $m_i(\mathbf{z}_i,\vec{\mu},\vec{\Gamma},\mathbf{\Lambda})$, we need to find the conditional distribution of $\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{Y}_i,\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)}$. Using Bayes rule, we have

$$p(\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{Y}_i,\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)}) = \frac{p(\mathbf{Y}_i|\boldsymbol{\gamma}_i^{\mathbf{z}_i},\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)})\,p(\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)})}{p(\mathbf{Y}_i|\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)})}$$

Note that all distributions involved in this expression are Gaussian. Let $\mathcal{N}_r(\mu,\Sigma)$ denote an r-variate Gaussian distribution with mean μ and covariance Σ. We have

$$\begin{aligned}
\mathbf{Y}_i|\boldsymbol{\gamma}_i^{\mathbf{z}_i},\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)} &\sim \mathcal{N}_{n_i}([\mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)} + \mathbf{B}_i\mathbf{\Theta}_{(t)}\boldsymbol{\mu}_{\mathbf{z}_i}^{(t)} + \mathbf{B}_i\mathbf{\Theta}_{(t)}\boldsymbol{\gamma}_{i(t)}^{\mathbf{z}_i}],[\sigma_{(t)}^2\nu_i^{(t)}I_{n_i}]) \\
\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)} &\sim \mathcal{N}_k(0,\mathbf{\Gamma}_{\mathbf{z}_i}^{(t)}) \\
\mathbf{Y}_i|\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)} &\sim \mathcal{N}_{n_i}([\mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)} + \mathbf{B}_i\mathbf{\Theta}_{(t)}\boldsymbol{\mu}_{\mathbf{z}_i}^{(t)}],[\sigma_{(t)}^2\nu_i^{(t)}I_{n_i} + \mathbf{B}_i\mathbf{\Theta}_{(t)}\mathbf{\Gamma}_{\mathbf{z}_i}^{(t)}\mathbf{\Theta}_{(t)}^\top\mathbf{B}_i^\top])
\end{aligned}$$

These simplifications lead to a multivariate Gaussian distribution

$$\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{Y}_i,\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\mathbf{\Gamma}}^{(t)},\mathbf{\Lambda}^{(t)} \sim \mathcal{N}_k(\hat{\gamma}_i^{\mathbf{z}_i},\hat{V}_i^{\mathbf{z}_i}),$$

with

$$\begin{aligned}
\hat{\gamma}_i^{\mathbf{z}_i} &= \left\{\nu_{(t)}^i\sigma_{(t)}^2\mathbf{\Gamma}_{\mathbf{z}_i}^{-1(t)} + \mathbf{\Theta}_{(t)}^\top\mathbf{B}_i^\top\mathbf{B}_i\mathbf{\Theta}_{(t)}\right\}^{-1}\mathbf{\Theta}_{(t)}^\top\mathbf{B}_i^\top\left\{\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i\mathbf{\Theta}_{(t)}\boldsymbol{\mu}_{\mathbf{z}_i}^{(t)}\right\} \\
\hat{V}_i^{\mathbf{z}_i} &= \left\{\mathbf{\Gamma}_{\mathbf{z}_i}^{-1(t)} + \frac{\mathbf{\Theta}_{(t)}^\top\mathbf{B}_i^\top\mathbf{B}_i\mathbf{\Theta}_{(t)}}{\nu_{(t)}^i\sigma_{(t)}^2}\right\}^{-1}
\end{aligned}$$

In the expression of $l_i(\vec{\mu},\vec{\Gamma},\mathbf{\Lambda})$, the random variable $\boldsymbol{\gamma}_i^{\mathbf{z}_i}$ only occurs in the terms $\left\|\mathbf{Y}_i - (\mathbf{B}_i\boldsymbol{\theta}_\mu + \mathbf{B}_i\mathbf{\Theta}\boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i\mathbf{\Theta}\boldsymbol{\gamma}_i^{\mathbf{z}_i})\right\|^2$ and $\boldsymbol{\gamma}_{i,\mathbf{z}_i}^\top\mathbf{\Gamma}_{\mathbf{z}_i}^{-1}\gamma_{i,\mathbf{z}_i}$. The other terms are left unchanged by the expectation. Consider the following identity that applies to any random vector U of dimension n.

$$\begin{aligned}
E(U^\top U) &= \text{trace}(E[U^\top U]) = E(\text{trace}[U^\top U]) = E(\text{trace}[UU^\top]) \\
&= \text{trace}(E[UU^\top]) = \hat{U}^\top\hat{U} + \text{trace}(\hat{V}_U),
\end{aligned}$$

where $\hat{U} = E(U)$, and $\hat{V}_U = Var(U)$. Using this, we get

$$\begin{aligned}
&E_{\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{Y}_i,\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\boldsymbol{\Gamma}}^{(t)},\boldsymbol{\Lambda}^{(t)}}\left\{\left\|\mathbf{Y}_i - (\mathbf{B}_i\boldsymbol{\theta}_\mu + \mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\gamma}_i^{\mathbf{z}_i})\right\|^2\right\} \\
&\qquad = \left\|\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\boldsymbol{\mu}_{\mathbf{z}_i}^{(t)} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\hat{\gamma}_i^{\mathbf{z}_i})\right\|^2 + \text{trace}\left[\mathbf{B}_i\boldsymbol{\Theta}_{(t)}\hat{V}_i^{\mathbf{z}_i}\boldsymbol{\Theta}_{(t)}^\top\mathbf{B}_i^\top\right],
\end{aligned}$$

and

$$E_{\boldsymbol{\gamma}_i^{\mathbf{z}_i}|\mathbf{Y}_i,\mathbf{z}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\boldsymbol{\Gamma}}^{(t)},\boldsymbol{\Lambda}^{(t)}}\left\{\boldsymbol{\gamma}_{i,\mathbf{z}_i}^\top\boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1}\boldsymbol{\gamma}_{i,\mathbf{z}_i}\right\} = \hat{\gamma}_{i,\mathbf{z}_i}^\top\boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1}\hat{\gamma}_{i,\mathbf{z}_i} + \text{trace}\left[\boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1}\hat{V}_i^{\mathbf{z}_i}\right],$$

which leads to the computation of $m_i(\mathbf{z}_i, \vec{\mu}, \vec{\Gamma}, \boldsymbol{\Lambda})$.
Next, we compute $E_{\mathbf{z}_i|\mathbf{Y}_i,\nu_i^{(t)},\vec{\boldsymbol{\mu}}^{(t)},\vec{\boldsymbol{\Gamma}}^{(t)},\boldsymbol{\Lambda}^{(t)}}\left\{m_i(\mathbf{z}_i, \vec{\mu}, \vec{\Gamma}, \boldsymbol{\Lambda})\right\}$. This requires finding the distribution of the discrete random variable $\left\{\mathbf{z}_i|\mathbf{Y}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}\right\}$. Let $\mathbf{e}_g$ be the G-dimensional vector whose components are all zero, except for the gth component which is set to 1. We have,

$$\begin{aligned}
&p(\mathbf{z}_i = \mathbf{e}_g|\mathbf{Y}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}) \\
&\quad = \frac{p(\mathbf{Y}_i|\mathbf{z}_i = \mathbf{e}_g, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)})\, p(\mathbf{z}_i = \mathbf{e}_g|\nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)})}{p(\mathbf{Y}_i|\nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)})} \\
&\quad = \frac{\pi_g^{(t)}\, p(\mathbf{Y}_i|\mathbf{z}_i = \mathbf{e}_g, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)})}{\sum_{h=1}^G \pi_h^{(t)}\, p(\mathbf{Y}_i|\mathbf{z}_i = \mathbf{e}_h, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)})}. \qquad (A.2)
\end{aligned}$$

The computation of the expression in (A.2) requires knowledge of the distribution of the random variable $\mathbf{Y}_i|\mathbf{z}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}$, which is a n_i-variate Gaussian random variable with

$$\begin{aligned}
\boldsymbol{E}_{i,\mathbf{z}_i} &= \mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)} + \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\boldsymbol{\mu}_{\mathbf{z}_i}^{(t)} \\
\boldsymbol{\Sigma}_{i,\mathbf{z}_i} &= \nu_{(t)}^i\sigma_{(t)}^2\left[I_{n_i} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\left(\nu_{(t)}^i\sigma_{(t)}^2\boldsymbol{\Gamma}_{\mathbf{z}_i}^{-1(t)} + \boldsymbol{\Theta}_{(t)}^\top\mathbf{B}_i^\top\mathbf{B}_i\boldsymbol{\Theta}_{(t)}\right)^{-1}\boldsymbol{\Theta}_{(t)}^\top\mathbf{B}_i^\top\right]
\end{aligned}$$

This distribution has been obtained by integrating out the individual random effects

$$p(\mathbf{Y}_i|\mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) = \int_{\boldsymbol{\gamma}_i} p(\mathbf{Y}_i, \boldsymbol{\gamma}_i|\mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda})\, d\boldsymbol{\gamma}_i,$$

where

$$\begin{aligned}
p(\mathbf{Y}_i, \boldsymbol{\gamma}_i|\mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) &= p(\mathbf{Y}_i|\boldsymbol{\gamma}_i, \mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda})\, p(\boldsymbol{\gamma}_i|\mathbf{z}_i, \nu_i, \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\Gamma}}, \boldsymbol{\Lambda}) \\
&= \mathcal{N}_{n_i}(\mathbf{B}_i\boldsymbol{\theta}_\mu + \mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\mu}_{\mathbf{z}_i} + \mathbf{B}_i\boldsymbol{\Theta}\boldsymbol{\gamma}_i^{\mathbf{z}_i}; \sigma^2\nu_i I_{n_i}) \times \mathcal{N}_k(\mathbf{0}; \boldsymbol{\Gamma}_{\mathbf{z}_i}).
\end{aligned}$$

Let $P_{ig}^{(t)}$ be the probability that the individual i belongs to group g. We have from Equation (A.2),

$$
\begin{aligned}
P_{ig}^{(t)} &= \pi_g^{(t)}\, F_{ig}^{(t)} \Big/ \sum_{h=1}^{G} (\pi_h^{(t)}\, F_{ih}^{(t)}) \quad \text{with} \\
F_{ig}^{(t)} &= \exp\left\{ -\frac{1}{2} \left(\mathbf{Y}_i - \boldsymbol{E}_{ig}^{(t)} \right)^{\mathsf{T}} \boldsymbol{\Sigma}_{ig}^{-1(t)} \left(\mathbf{Y}_i - \boldsymbol{E}_{ig}^{(t)} \right) \right\} \Big/ \left\{ (2\pi)^{n_i/2} |\boldsymbol{\Sigma}_{ig}^{(t)}|^{1/2} \right\} \qquad g = 1, \ldots, G.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
& E_{\mathbf{z}_i | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}} \left\{ m_i(\mathbf{z}_i, \vec{\mu}, \vec{\Gamma}, \boldsymbol{\Lambda}) \right\} \\
= & \sum_{g=1}^{G} p\left(\mathbf{z}_i = \mathbf{e}_g | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)} \right) \times m_i \left(\mathbf{z}_i = \mathbf{e}_g, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)} \right) \\
= & \sum_{g=1}^{G} P_{ig}^{(t)} \times m_i \left(\mathbf{z}_i = \mathbf{e}_g, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)} \right).
\end{aligned}
$$

Thus, from Equation (A.1), we have:

$$
\begin{aligned}
Q(\boldsymbol{\Pi} | \boldsymbol{\Pi}^{(t)}) &= \sum_{i=1}^{N} E_{\mathbf{z}_i | \mathbf{Y}_i, \nu_i^{(t)}, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)}} \left\{ m_i(\mathbf{z}_i, \vec{\mu}, \vec{\Gamma}, \boldsymbol{\Lambda}) \right\} + \mathcal{H} \\
&= \sum_{i=1}^{N} \sum_{g=1}^{G} P_{ig}^{(t)} \times m_i \left(\mathbf{z}_i = \mathbf{e}_g, \vec{\boldsymbol{\mu}}^{(t)}, \vec{\boldsymbol{\Gamma}}^{(t)}, \boldsymbol{\Lambda}^{(t)} \right) + \mathcal{H}.
\end{aligned}
$$

Finally, after the expectation step, the expression of the function $Q(\boldsymbol{\Pi}|\boldsymbol{\Pi}^{(t)})$ where all the parameters are at their t^{th} updated value is given by:

$$
\begin{aligned}
Q(\boldsymbol{\Pi}|\boldsymbol{\Pi}^{(t)}) \quad = \quad & \sum_{i=1}^{N} \sum_{g=1}^{G} P_{ig} \times \left\{ \begin{array}{l}
-\frac{n_i}{2} \log(\nu_i \sigma^2) - \frac{1}{2} \log(\left|\boldsymbol{\Gamma}_g\right|) + \log(\pi_g) \\
-\frac{1}{2\nu_i\sigma^2} \{ \left\| \mathbf{Y}_i - \mathbf{B}_i \boldsymbol{\theta}_\mu - \mathbf{B}_i \boldsymbol{\Theta} \boldsymbol{\mu}_g - \mathbf{B}_i \boldsymbol{\Theta} \hat{\gamma}_i^g) \right\|^2 \} \\
+\frac{1}{2\nu_i\sigma^2} \{ \text{ trace } \left[\mathbf{B}_i \boldsymbol{\Theta} \hat{V}_i^g \boldsymbol{\Theta}^{\mathsf{T}} \mathbf{B}_i^{\mathsf{T}} \right] \} \\
-\frac{1}{2} \{ \hat{\boldsymbol{\gamma}}_{ig} \boldsymbol{\Gamma}_g^{-1} \hat{\boldsymbol{\gamma}}_{ig} + \text{trace} \left[\boldsymbol{\Gamma}_g^{-1} \hat{V}_i^g \right] \} \\
+\frac{\nu_o}{2} \log(\frac{\nu_o}{2}) - \log[\Gamma(\frac{\nu_o}{2})] - (1 + \frac{\nu_o}{2}) \log(\nu_i) - \frac{\nu_o}{2\nu_i}
\end{array} \right\} \\
& + \sum_{g=1}^{G} \left\{ \begin{array}{l}
-\frac{1}{2} \log(\left|\boldsymbol{\Gamma}_\mu\right|) - \frac{1}{2} \boldsymbol{\mu}_g^T \boldsymbol{\Gamma}_\mu^{-1} \boldsymbol{\mu}_g + \frac{m}{2} \log(\left|(m-k-1)\mathbf{D}\right|) \\
-\frac{(m+k+1)}{2} \log(\left|\boldsymbol{\Gamma}_g\right|) - \frac{(m-k-1)}{2} \text{trace}[\mathbf{D}\boldsymbol{\Gamma}_g^{-1}]
\end{array} \right\} \\
& + \quad \frac{km}{2} \log(m-k-1) - \frac{(m+k+1)}{2} \log(\left|\boldsymbol{\Gamma}_\mu\right|) - \frac{(m-k-1)}{2} \text{trace}[\boldsymbol{\Gamma}_\mu^{-1}] \\
& + \sum_{j=1}^{k} \left\{ \; +\frac{m}{2} \log(\frac{m}{2}) - \log[\Gamma(\frac{m}{2})] - (1 + \frac{m}{2}) \log(d_{jj}) - \frac{m}{2d_{jj}} \; \right\} \\
& + \quad \alpha_\sigma \log(\beta_\sigma) - \log[\Gamma(\alpha_\sigma)] - (\alpha_\sigma + 1) \log(\sigma^2) - \frac{\beta_\sigma}{\sigma^2} \\
& + \quad -\log[B(a_1, ..., a_G)] + \textstyle\sum_{g=1}^{G} (a_g - 1) \log(\pi_g) \quad + \mathcal{C}
\end{aligned}
$$

B The EM updating equations for the Bayesian lasso FCM

$$\begin{cases} \boldsymbol{\Gamma}_g^{(t+1)} = \left[\left(\sum_{i=1}^N P_{ig}\right) + (m+k+1)\right]^{-1} \left[\left(\sum_{i=1}^N P_{ig}(\hat{\boldsymbol{\gamma}}_{ig}\hat{\boldsymbol{\gamma}}_{ig}^\mathsf{T} + \hat{V}_i^g)\right) + (m-k-1)\mathbf{D}^{(t)}\right] \\ \pi_g^{(t+1)} = \frac{\left(\sum_{i=1}^N P_{ig}\right)+(a_g-1)}{N+\left(\sum_{g=1}^G a_g\right)-G} \quad (0 \leq \pi_g^{(t+1)} \leq 1) \end{cases}$$

$$\begin{cases} \boldsymbol{\theta}_\mu^{(t+1)} = \left[\sum_{i=1}^N \left(\mathbf{B}_i^\mathsf{T}\mathbf{B}_i\right) \nu_i^{-1(t)}\right]^{-1} \left[\sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}}\mathbf{B}_i^\mathsf{T} \left(\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\hat{\boldsymbol{\gamma}}_i^g\right)\right] \\ \boldsymbol{\Theta}_j^{(t+1)} = \left\{\sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[(\boldsymbol{\alpha}_{ig})_j^2 + (\hat{V}_i^g)_{jj}\right] \left(\mathbf{B}_i^\mathsf{T}\mathbf{B}_i\right)\right\}^{-1} \{\Omega_1 - \Omega_2\} \quad ; \quad \text{for} \quad j = 1, ..., k. \\ \Omega_1 = \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[(\boldsymbol{\alpha}_{ig})_j\mathbf{B}_i^\mathsf{T} \left(\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)}\right)\right]; \quad (\boldsymbol{\alpha}_{ig})_j = (\boldsymbol{\mu}_g + \hat{\boldsymbol{\gamma}}_i^g)_j \\ \Omega_2 = \sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}} \left[\sum_{h\neq j}^k \left((\boldsymbol{\alpha}_{ig})_j(\boldsymbol{\alpha}_{ig})_h + (\hat{V}_{ig})_{hj}\right) \left(\mathbf{B}_i^\mathsf{T}\mathbf{B}_i\right) \boldsymbol{\Theta}_h\right] \end{cases}$$

$$\begin{cases} \sigma^2_{(t+1)} = \frac{\frac{1}{2}\left\{\sum_{i=1}^N \sum_{g=1}^G \frac{P_{ig}}{\nu_i^{(t)}}\left[\left\|\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\hat{\boldsymbol{\gamma}}_i^g\right\|^2 + trace\left(\mathbf{B}_i\boldsymbol{\Theta}_{(t)}\hat{V}_{ig}\boldsymbol{\Theta}_{(t)}^\mathsf{T}\mathbf{B}_i^\mathsf{T}\right)\right]\right\} + \beta_\sigma}{\frac{1}{2}\left[\sum_{i=1}^N n_i\right] + (\alpha_\sigma + 1)} \\ \nu_0^{(t+1)} = \frac{b_{\nu_0}+1+\sqrt{2b_{\nu_0}+1}}{b_{\nu_0}} \quad \text{with} \quad b_{\nu_0} = \exp\left(\frac{1}{N}\sum_{i=1}^N (\log \nu_i^{(t)} + 1/\nu_i^{(t)}) - 1\right) \\ \nu_i^{(t+1)} = \frac{\left\{\nu_0^{(t)} + \sum_{g=1}^G \frac{\hat{P}_{ig}}{\sigma^2_{(t)}}\left[\left\|\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\boldsymbol{\mu}_g^{(t)} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\hat{\boldsymbol{\gamma}}_i^g\right\|^2 + trace\left(\mathbf{B}_i\boldsymbol{\Theta}_{(t)}\hat{V}_{ig}\boldsymbol{\Theta}_{(t)}^\mathsf{T}\mathbf{B}_i^\mathsf{T}\right)\right]\right\}}{n_i + 2 + \nu_0^{(t)}} \\ d_{jj}^{(t+1)} = \frac{1}{2}\left[\frac{(mG-m-2) + \sqrt{(mG-m-2)^2 + 4m\times(m-k-1)\times\sum_{g=1}^G\left\{\boldsymbol{\Gamma}_g^{-1(t)}\right\}_{jj}}}{(m-k-1)\times\sum_{g=1}^G\left\{\boldsymbol{\Gamma}_g^{-1(t)}\right\}_{jj}}\right] \quad (j = 1, \ldots, k) \\ \left\{\boldsymbol{\Gamma}_\mu^{(t+1)}\right\}_{jj} = \frac{1}{m+k+1+G}\left[\left\{\sum_{g=1}^G \boldsymbol{\mu}_{g(t)}\boldsymbol{\mu}_{g(t)}^\mathsf{T}\right\}_{jj} + (m-k-1)\right]. \end{cases}$$

The results of the M-step for the cluster means $\boldsymbol{\mu}_g$ are presented below for each of the two options of the $D^{ist}(\cdot,\cdot)$ function in the second penalty term from equation (6).

$$\begin{cases} A_1 & = \left[\sum_{i=1}^{N} \frac{P_{ig}}{\nu_i^{(t)}\sigma^2_{(t)}} \left(\left(\mathbf{Y}_i - \mathbf{B}_i\boldsymbol{\theta}_\mu^{(t)} - \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\hat{\gamma}_i^g\right)^\top \mathbf{B}_i\boldsymbol{\Theta}_{(t)}\right)_j\right] \\ & \quad + \left[\sum_{i=1}^{N} \frac{-P_{ig}}{2\nu_i^{(t)}\sigma^2_{(t)}} \sum_{r\neq j}^{k} \mu_g^{r(t)} \left(\boldsymbol{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \boldsymbol{\Theta}_{(t)}\right)_{jr}\right] \\ B_1 & = \left[\sum_{i=1}^{N} \left\{\frac{P_{ig}}{\nu_i^{(t)}\sigma^2_{(t)}} \left(\boldsymbol{\Theta}_{(t)}^\top \mathbf{B}_i^\top \mathbf{B}_i \boldsymbol{\Theta}_{(t)}\right)_{jj}\right\} + \left\{\boldsymbol{\Gamma}_\mu^{-1(t)}\right\}_{jj}\right] \\ \mu_g^{j(t+1)} & = \left[\frac{A_1 - \lambda_1\left\{sign(\mu_g^{j(t)})\right\} + \left\{\lambda_2 \sum_{h\neq g}^{G} \frac{\mu_h^{j(t)}}{D^{ist}(\boldsymbol{\mu}_g,\boldsymbol{\mu}_h)}\right\}}{B_1 + \left\{\lambda_2 \sum_{h\neq g}^{G} \frac{1}{D^{ist}(\boldsymbol{\mu}_g,\boldsymbol{\mu}_h)}\right\}}\right] \qquad \text{for the } L_2 \text{ norm distance} \\ \mu_g^{j(t+1)} & = \left[\frac{A_1 - \lambda_1\left\{sign(\mu_g^{j(t)})\right\} - \lambda_2\left\{\sum_{h=1}^{G} sign(\mu_g^{j(t)} - \mu_h^{j(t)})\right\}}{B_1}\right] \qquad \text{for the } L_1 \text{ norm distance.} \end{cases}$$

Note that $sign(x)$ equals -1 if $x < 0$, equals 1 if $x > 0$ and equals 0 if $x = 0$. For identifiability reasons, we assume the matrices Γ_μ and D to be diagonal. Also note that in general the matrix Θ output by the procedure will not necessarily be orthonormal. Therefore, we transform the output matrix into an orthonormal matrix with the Gram-Schmidt algorithm Golub and Van Loan [1996].