Wavelet-based clustering for mixed-effects functional models in high dimension.

M. Giacofci[†], S. Lambert-Lacroix[◦], G. Marot[◇,*,▽] F. Picard[⋆]

[†] *Laboratoire LJK, BP 53, Université de Grenoble et CNRS, 38041 Grenoble cedex 9, France*
[◦] *UJF-Grenoble 1/CNRS/UPMF/TIMC-IMAG UMR 5525, Grenoble, F-38041, France*
[⋆] *LBBE, UMR CNRS 5558 Université Lyon 1, F-69622, Villeurbanne, France*
[◇] *Projet BAMBOO, INRIA Rhône-Alpes, F-38330 Montbonnot Saint-Martin, France.*
[*] *Biostatistics, EA 2694, UDSL, Université Lille Nord de France.*
[▽] *MODAL, INRIA Lille Nord Europe,F-59650 Villeneuve d'Ascq, France.*

Abstract

We propose a method for high dimensional curve clustering in the presence of inter-individual variability. Curve clustering has longly been studied especially using splines to account for functional random effects. However splines are not appropriate when dealing with high-dimensional data and can not be used to model irregular curves such as peak-like data. Our method is based on a wavelet decomposition of the signal for both fixed and random-effects. We propose an efficient dimension reduction step based on wavelet thresholding adapted to multiple curves and using an appropriate structure for the random effect variance, we ensure that both fixed and random effects lie in the same functional space even when dealing with irregular functions that belong to Besov spaces. In the wavelet domain our model resumes to a linear mixed-effects model that can be used for a model-based clustering algorithm and for which we develop an EM-algorithm for maximum likelihood estimation. The properties of the overall procedure are validated by an extensive simulation study. Then we illustrate our method on mass spectrometry data and we propose an original application of functional data analysis on microarray CGH data. Our procedure is available through the R package `curvclust` which is the first publicly available package that performs curve clustering with random effects in the high dimensional framework.

Keywords: Clustering; Functional data; Mixed models; Wavelets.

Contact: `madison.giacofci@imag.fr`

1. Introduction

Functional data analysis has gained increased attention in the past years, in particular in high-throughput biology with the use of mass spectrometry. This method is used to characterize the protein content of biological samples by separating compounds according to their mass to charge ratio (m/z). Among different technologies Matrix Assisted Laser Desorption and Ionization, Time-Of-Flight (MALDI-TOF) mass spectrometry is one the most used and has become standard to improve proteomic profiling of diseases as well as clinical diagnosis.

Dedicated methods have been developed to analyze such data for differential analysis, supervised classification and clustering (Hilario et al., 2006). Up to now the functional setting has mostly been developed for differential analysis (Morris et al., 2008). One central element is the modeling of the inter-individual variability by using functional random effects, since subject-specific fluctuations are known to be the largest source of variability in mass-spec data (Eckel-Passow et al., 2009). In this paper we focus on the non supervised task which consists in finding groups of individuals whose proteomic landscape is similar. Surprisingly the clustering task received less attention, and is mainly based on hierarchical clustering on the set of peaks detected across spectra (Bensmail et al., 2005; Morris et al., 2010). However such method is known to depend heavily on the peak detection method and has the strong dis-advantage to neglect the inter-individual variability whereas this information should be central for subgroup discovery. Thus our main focus in this paper is modelling and clustering curves of this type in a functional mixed model framework.

When dealing with curve clustering in the presence of individual variability, a pioneer work is based on a spline decomposition of the signal (James and Sugar, 2003) which resumes to a linear mixed effect model on which clustering and low-dimensional representation can be performed. However splines show two main drawbacks: i) they are inappropriate when dealing with functions that show peaks and irregularities, ii) they require heavy computational efforts and so are not adapted to high dimensional data. On the contrary, wavelet representations appear to be a natural framework to consider such irregularities through the sequence space of (usually sparse) Besov representation. Recent works have been done about estimation and inference in the functional mixed effects framework based on a wavelet decomposition approach. A fully Bayesian version has been proposed by Morris and Carroll (2006), with non-parametric estimates of fixed and random effects as well as between and within-curve covariance matrix estimates to accomodate a wide variety of correlation structures. In addition, Antoniadis and Sapatinas (2007b) propose a study of both estimation and inference in a frequentist framework.

In this paper we use a wavelet representation for both fixed and random effects to perform model-based clustering. Such strategy has been considered by Antoniadis et al. (2008) and by Ray and Mallick (2006) without random effects for image clustering and for the analysis of time course experiments respectively. We use a similar approach and we extend it by adding functional random effects. Inter-individual variability in the wavelet domain is modeled using results of Antoniadis and Sapatinas (2007b) but

accomodates a broader range of correlation structure. In particular we allow within curve correlation to vary over groups and positions. Then we propose a two-step procedure which involves a dimension reduction step and a clustering step based on the EM-algorithm. We also propose a model-selection criterion that accounts for the inter-individual variability, and we define a rigorous simulation framework for curve clustering. Our method is implemented within the `R` package `curvclust` which is the first available software dedicated to this task. In a first application, we illustrate our method on the mass spectrometry data first published in Petricoin et al. (2002).

Then our last contribution is to extend the use of functional models to another type of high throughput data which are Comparative Genomic Hybridization (CGH) data. The CGH array technology is used to map copy number imbalances between genomes by hybridizing differentially labeled genomic DNAs on a chip. Fluorescence ratios are usually analyzed using change-point models to detect segments that correspond to homogeneous regions on the genome in terms of copy number. Clustering patients based on their CGH profiles is very promising and has been successfully used to identify molecular subtypes of cancer. However clustering CGH profiles based on a segmentation has the same drawbacks that clustering mass spectra based on detected peaks: results depends on the segmentation methods. Moreover the inter-individual variability has never been investigated in this type of data, whereas it is likely to represent an important part of the variability of the data especially for cancer profiles. We use the breast cancer data of Fridlyand et al. (2006) that have already been analysed for non-supervised clustering by Van Wieringen et al. (2008). We show the interest of functional random effects for these type of and we discuss the impacts in terms of analysis and design for copy number studies.

2. Functional Clustering modeling using wavelets

2.1 Presentation of the model

We observe n curves $Y_i(t)$ over M equally spaced time points $\mathbf{t} = (t_1, \ldots, t_M)$ in $[0, 1]$, with $M = 2^J$ for some integer J and we model these data by the linear functional model of the form:

$$Y_i(t) = \mu_i(t) + E_i(t), \quad E_i(t) \sim \mathcal{N}(0, \sigma_E^2). \tag{1}$$

In the following we will use notation $\mathbf{Y}_i(\mathbf{t}) = [Y_i(t_1), \ldots, Y_i(t_M)]$. In the functional clustering setting we suppose that individuals are spread among L unknown clusters of prior size π_ℓ, $\ell = 1, \ldots, L$, and we denote by $\zeta_{i\ell}$ the indicator variable that equals 1 if the ith individual is in the ℓth group. Given $\{\zeta_{i\ell} = 1\}$, model (1) becomes

$$Y_i(t) = \mu_\ell(t) + E_i(t), \tag{2}$$

where $\mu_\ell(t)$ is the principal functional fixed effect that characterizes cluster ℓ. To handle subject-specific random deviations from the cluster average curve we introduce random functions $U_i(t)$ that are modelled as centered Gaussian processes not necessarily stationary but independent from $E_i(t)$. Then given $\{\zeta_{i\ell} = 1\}$, model 2 becomes

$$Y_i(t) = \mu_\ell(t) + U_i(t) + E_i(t), \quad U_i(t) \sim \mathcal{N}(0, K_\ell(s, t)) \tag{3}$$

Once defined in the functional domain, the classical approach is to convert the original infinite-dimensional clustering problem into a finite-dimensional problem using a functional basis representation of the model. At this step James and Sugar (2003) propose a spline-based representation of model (3) with individuals observed at sparse sets of time points like in longitudinal data. Our procedure is more adapted to high dimensional data thanks to the computational efficiency of wavelets, unlike splines that require matrix inversions whose complexity increases with the density of the design. Moreover, as we will see below, the wavelet representation allows us to account for a wider range of functional shapes than splines, thanks to their connection with Besov spaces.

Using a wavelet representation of this model allows us to characterize different types of smoothness conditions assumed on the response curves $Y_i(t)$ by the mean of their wavelet coefficients. Moreover wavelet representations are sparse for a wide variety of functional spaces, which is crucial when dealing with high dimensional data. This property will be central while performing dimension reduction. Briefly, we are working with an orthonormal wavelet basis

$$\{\phi_{j_0k}(t), k = 0, 1, \dots 2^{j_0} - 1; \psi_{jk}(t), j \geq j_0, k = 0, \dots, 2^j - 1\}$$

generated from a father wavelet ϕ and a mother wavelet ψ of regularity r, ($r \geq 0$). In this basis the response curve $Y_i(t)$ has the following decomposition:

$$Y_i(t) = \sum_{k=0}^{2^{j_0}-1} c^*_{i,j_0k}\phi_{j_0k}(t) + \sum_{j\geq j_0}\sum_{k=0}^{2^j-1} d^*_{i,jk}\psi_{jk}(t).$$

In practice we use the Discrete Wavelet Transform (DWT) which can be performed thanks to Mallat's fast algorithm with $\mathcal{O}(M)$ operations only. We denote by $\mathbf{W}$ the $(M \times M)$-matrix containing filters of the chosen wavelet basis. The resulting scaling and wavelet coefficients $\mathbf{c}_i = [c_{i,j_0k}]_{(k)}$ and $\mathbf{d}_i = [d_{i,jk}]_{(jk)}$ of the individual curves are empirical coefficients. They are related to their continuous counterparts c^*_{i,j_0k} and $d^*_{i,jk}$ by: $c_{i,j_0k} \approx \sqrt{M}c^*_{i,j_0k}$ and $d_{i,jk} \approx \sqrt{M}d^*_{i,jk}$. Moreover, without loss in generality we assume that $j_0 = 0$. When applying the DWT to model (3) we have

$$\mathbf{W}\mathbf{Y}_i(\mathbf{t}) = \mathbf{W}\boldsymbol{\mu}_\ell(\mathbf{t}) + \mathbf{W}\mathbf{U}_i(\mathbf{t}) + \mathbf{W}\mathbf{E}_i,$$

which resumes to a linear mixed-effect model in the coefficient domains such that

$$\begin{aligned} \mathbf{c}_i &= \boldsymbol{\alpha}_\ell + \boldsymbol{\nu}_i + \boldsymbol{\varepsilon}_i \\ \mathbf{d}_i &= \boldsymbol{\beta}_\ell + \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i. \end{aligned}$$

$(\boldsymbol{\alpha}_\ell, \boldsymbol{\beta}_\ell)$ stand for the scaling and wavelet coefficients of the fixed average curve $\boldsymbol{\mu}_\ell(\mathbf{t})$, and $(\boldsymbol{\nu}_i, \boldsymbol{\theta}_i)$ are the scaling and wavelet random coefficients of Gaussian process $\mathbf{U}_i(\mathbf{t})$ such that

$$\begin{bmatrix} \boldsymbol{\nu} \\ \boldsymbol{\theta} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \mathbf{G}_{\boldsymbol{\nu}} & 0 \\ 0 & \mathbf{G}_{\boldsymbol{\theta}} \end{bmatrix}\right),$$

and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}) \perp (\boldsymbol{\nu}, \boldsymbol{\theta})'$, with $\sigma_\varepsilon^2 = \sigma_E^2 / M$. Note that the model we present could be extended to more general functional mixed models

$$\mathbf{Y}_i(\mathbf{t}) = \mathbf{X}_i \boldsymbol{\mu}_\ell(\mathbf{t}) + \mathbf{Z}_i \mathbf{U}_i(\mathbf{t}) + \mathbf{E}_i,$$

hence our method can be used for the clustering of functional data based on linear fixed and random functional effects. For the sake of simplicity the derivation of our method is restricted to the case $\mathbf{X} = \mathbf{I}$, $\mathbf{Z} = \mathbf{I}$.

2.2 Besov spaces and specification of the variance of random effects

The strength of the wavelet representation is that it allows us to handle very diverse shapes of curves among which curves with irregularities that lie in particular Besov spaces. Besov spaces consist of functions that have a specific degree of smoothness. Roughly speaking, for a Besov space $B^s_{p,q}[0,1]$, parameter s indicates the number of function's derivatives, where their existence is required in a L^p-sense, q allowing finer control of the function's regularity. For a detailed study of Besov spaces, we refer to Donoho and Johnstone (1998). When dealing with functional mixed-effect models, the difficulty is that if the fixed-effect curve $\boldsymbol{\mu}_\ell(\mathbf{t})$ is supposed to belong to some Besov space, then the subject-specific deviations arising from the random functions $\mathbf{U}_i$ should be controled so that $\mathbf{U}_i$ belongs to the same functional space. As proposed by Abramovich et al. (1998) and more specifically by Antoniadis and Sapatinas (2007b) in the context of functional mixed-models, this goal is achieved by controlling the exponential decrease of the variances of the random wavelet coefficients such that:

$$\mathbf{G}_{\boldsymbol{\theta}} = \underset{jk}{\text{Diag}} \left(2^{-j\eta} \gamma_\theta^2\right), \ \forall j \in \{j_0, \ldots, J\}, \ k \in \{0, \ldots, 2^j - 1\}.$$

This control requires the introduction of parameter η which is associated with the regularity of process $\mathbf{U}_i(\mathbf{t})$. Abramovich et al. (1998) state that given a mother wavelet ψ of regularity r, where $\max(0, \frac{1}{p} - \frac{1}{2}) < s < r$ and given that $\mu_\ell(t) \in B^s_{p,q}[0,1]$, then:

$$U_i(t) \in B^s_{p,q}[0,1] \text{ a.s.} \quad \Longleftrightarrow \quad \begin{cases} s + \frac{1}{2} - \frac{\eta}{2} = 0 & \text{if } 1 \leq p < \infty \text{ and } q = \infty, \\ s + \frac{1}{2} - \frac{\eta}{2} < 0 & \text{otherwise.} \end{cases}$$

At last it can be necessary to allow variance γ_θ^2 to depend on both scale and position ($\gamma^2_{\theta,jk}$) as pointed by Morris and Carroll (2006). Similarly the model can be enriched by considering a cluster-specific random effect variance $\gamma^2_{\theta,\ell}$ or $\gamma^2_{\theta,\ell jk}$. This modelling can be very powerful to consider different types of random functions $\mathbf{U}_i$.

2.3 Dimensionality reduction

Wavelet representations are sparse for a wide class of functional spaces which makes their use very efficient when dealing with high dimensional data. In the case of a single curve, shrinkage estimation and hard thresholding have been developed to set to zero

the wavelet coefficients whose absolute value is below the threshold $\widehat{\sigma}\sqrt{2\log M}$. The estimator $\widehat{\sigma}$ is usually given by the median absolute deviation ($\widehat{\sigma}_{\text{MAD}}$) of empirical wavelet coefficients at the finest resolution level divided by 0.6745. In this case, thresholding has the double advantage to reduce dimensionality and to ensure good reconstruction properties. Then in this step σ_ε is estimated with the average of the n robust estimates. In the framework of curve clustering, our goal is to reduce the dimensionality of the problem to handle heavy datasets and not to find the optimal reconstruction rule. With this in mind we follow the strategy proposed by Antoniadis et al. (2008) and we first perform an individual denoising in order to keep coefficients which contain individual-specific information. This is done by applying non-linear wavelet hard thresholding of the coefficients $\mathbf{d}_i$ via an universal threshold as described in Donoho and Johnstone (1994).

Then caution should be taken for the estimation of the noise measurement error. Firt we consider the average of the n robust estimates since we have many observed curves. Then the variance of the observations is $\mathbb{V}(d_{ijk}) = 2^{-j\eta}\gamma_\theta^2 + \sigma_\varepsilon^2$ due to the mixed model structure. Consequently its estimation would require estimates of both parameters σ_ε^2 and γ_θ^2. This can be easily done when the individual labels are known. Then provided that estimation of parameters σ_ε^2 and γ_θ^2 are available, the same threshold could be simply extended with the level dependent variance. Otherwise, this estimation is a difficult task when individual labels are unknown since it leads to estimate variance from samples with different and unknown means. Thereby we suggest to use $\widehat{\sigma}_{\text{MAD}}$ in our procedure even in the framework of mixed models. In this case $\widehat{\sigma}_{\text{MAD}}$ estimates the global variance at the finest resolution level which is equal to

$$\mathbb{V}(d_{iJk}) = 2^{-J\eta}\gamma_\theta^2 + \sigma_\varepsilon^2 \simeq \sigma_\varepsilon^2. \tag{4}$$

Note that using a level dependent thresholding that considers random effects would lead to greater variance estimate and hence to a greater dimensionality reduction. However this estimation is not possible in the non-supervised setting since the group-spectific means are unknown *a priori*. Moreover simulations showed that the difference was negligible (not shown). Finally we take the union set of wavelet coefficients that survived thresholding (Antoniadis et al., 2008). This method removes coefficients that are zeros for all individuals, and hence which are non informative regarding to the clustering goal. Note that we do not use the second reduction step proposed by Antoniadis et al. (2008) that consists in using a truncation procedure based on a Neyman test to increase the sparsity (Fan, 1996). In this step they test whether, for each wavelet coefficient in the representative union, its expectation across the curves remains constant against the assumption that its expected behavior differs among curves. It makes sense since in their case the curves are pixel-wise intensity curves in image segmentation. This image structure induces a coherence between adjacent pixels that makes consecutive differences sparse (see Antoniadis et al. (2008) for more details).

3. Parameter estimation and model selection

3.1 An EM algorithm for Maximum Likelihood estimation

Once projected in the wavelet domain, the clustering model resumes to a standard clustering model with additional random effects whose variance is of particular form. Thus parameters are estimated by maximum likelihood using the EM algorithm. Both label variables $\boldsymbol{\zeta}$ and random effects $(\boldsymbol{\nu}, \boldsymbol{\theta})$ are unobserved and the complete data log-likelihood can be written such that:

$$\begin{aligned}\log \mathcal{L}\left(\mathbf{c}, \mathbf{d}, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta}; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{G}, \sigma_\varepsilon^2\right) &= \log \mathcal{L}\left(\mathbf{c}, \mathbf{d} | \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta}; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\varepsilon^2\right) \\ &+ \log \mathcal{L}\left(\boldsymbol{\nu}, \boldsymbol{\theta} | \boldsymbol{\zeta}; \mathbf{G}\right) \\ &+ \log \mathcal{L}\left(\boldsymbol{\zeta}; \boldsymbol{\pi}\right).\end{aligned}$$

This likelihood can be easily computed thanks to the properties of mixed linear models such that:

$$\left[\begin{array}{c}\mathbf{c}_i \\ \mathbf{d}_i\end{array}\right] \Bigg| \left[\begin{array}{c}\boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i\end{array}\right], \{\zeta_{i\ell} = 1\} \sim \mathcal{N}\left(\left[\begin{array}{c}\boldsymbol{\alpha}_\ell + \boldsymbol{\nu}_i \\ \boldsymbol{\beta}_\ell + \boldsymbol{\theta}_i\end{array}\right], \sigma_\varepsilon^2 \mathbf{I}\right).$$

The EM algorithm provides the *posterior* probability of membership to cluster ℓ, $\tau_{i\ell}$ which is updated such that:

$$\tau_{i\ell}^{[h+1]} = \frac{\pi_\ell^{[h]} f\left(\mathbf{c}_i, \mathbf{d}_i; \boldsymbol{\alpha}_\ell^{[h]}, \boldsymbol{\beta}_\ell^{[h]}, \mathbf{G}^{[h]} + \sigma_\varepsilon^{2[h]} \mathbf{I}\right)}{\sum_p \pi_p^{[h]} f\left(\mathbf{c}_i, \mathbf{d}_i; \boldsymbol{\alpha}_p^{[h]}, \boldsymbol{\beta}_p^{[h]}, \mathbf{G}^{[h]} + \sigma_\varepsilon^{2[h]} \mathbf{I}\right)},$$

with $f()$ the probability density function of the Gaussian distribution. Moreover, using Hendersons trick we get the linear prediction of the random effects such that:

$$\begin{aligned}\widehat{\boldsymbol{\nu}}_{i\ell}^{[h+1]} &= \left(\mathbf{c}_i - \boldsymbol{\alpha}_\ell^{[h]}\right) / \left(1 + \lambda_\nu^{[h]}\right) \\ \widehat{\boldsymbol{\theta}}_{i\ell}^{[h+1]} &= \left(\mathbf{d}_i - \boldsymbol{\beta}_\ell^{[h]}\right) / \left(1 + 2^{j\eta} \lambda_\theta^{[h]}\right)\end{aligned}$$

with $(\lambda_\nu, \lambda_\theta) = (\sigma_\varepsilon^2/\gamma_\nu^2, \sigma_\varepsilon^2/\gamma_\theta^2)$. As for the maximization part, it provides the estimators of the mean curve coefficients

$$\begin{aligned}\boldsymbol{\alpha}_\ell^{[h+1]} &= \sum_{i=1}^{n} \tau_{i\ell}^{[h]} \left(\mathbf{c}_i - \widehat{\boldsymbol{\nu}}_{i\ell}^{[h]}\right) / N_\ell^{[h]}, \\ \boldsymbol{\beta}_\ell^{[h+1]} &= \sum_{i=1}^{n} \tau_{i\ell}^{[h]} \left(\mathbf{d}_i - \widehat{\boldsymbol{\theta}}_{i\ell}^{[h]}\right) / N_\ell^{[h]},\end{aligned}$$

with $N_\ell = \sum_i \tau_{i\ell}$. Moreover, the EM algorithm provides a ML estimator of the random effect variance such that:

$$\begin{aligned}
\gamma_\theta^{2[h+1]} &= \frac{1}{n(M-1)} \sum_{ijk\ell} 2^{j\eta} \tau_{i\ell}^{[h]} \left(\widehat{\theta}_{ijk\ell}^{2\ [h]} + \frac{\sigma_\varepsilon^{2[h]}}{1 + 2^{j\eta} \lambda_\theta^{[h]}} \right), \\
\gamma_\nu^{2[h+1]} &= \frac{1}{n} \sum_{i\ell} \left(\widehat{\nu}_{i00\ell}^{2\ [h]} + \frac{\sigma_\varepsilon^{2[h]}}{1 + \lambda_\nu^{[h]}} \right).
\end{aligned}$$

As last point, we mention that η can be estimated by maximization of the likelihood using the golden search section algorithm Kiefer (1953) rather than gridded optimization, and the EM algorithm can be speeded-up using the vector-ϵ algorithm as proposed by Kurodaa and Sakakiharab (2006).

3.2 Choosing the number of clusters using a BIC

We propose to choose the number of clusters using the framework of penalized likelihoods. In the following we use notations $\mathbf{m}_L[\gamma^2], \mathbf{m}_L[\gamma_\ell^2]$ for clustering models with L groups with constant and heterogeneous variances respectively. We first use the Bayesian Information Criterion and we select the dimension that maximizes

$$\mathrm{BIC}(\mathbf{m}_L[\gamma^2]) = \log \mathcal{L}\left(\mathbf{c}, \mathbf{d}; \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{G}}, \widehat{\sigma}_\varepsilon^2, \mathbf{m}_L[\gamma^2]\right) - \frac{|\mathbf{m}_L[\gamma^2]|}{2} \times \log(N).$$

This classical criterion is a penalized version of the observed-data log-likelihood where $|\mathbf{m}_L[\gamma^2]|$ is the number of free parameters of a model with L clusters, with $|\mathbf{m}_L[\gamma^2]| = |\boldsymbol{\alpha}| + |\boldsymbol{\beta}| + |\mathbf{G}| + |\boldsymbol{\pi}| - 1 + |\sigma_\varepsilon^2| = (M+1)L + |\mathbf{G}|$, the dimension of $\mathbf{G}$ depending on the variance structure of the random effects.

When considering mixed models, it is likely that the prediction of the random effects provides information regarding the number of clusters to select. In order to use information from hidden variables we propose to derive an Integrated Classification Likelihood criterion in the spirit of Biernacki et al. (2000). The ICL criterion is based on the integrated likelihood of the complete data:

$$\log \mathcal{L}(\mathbf{c}, \mathbf{d}, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta} | \mathbf{m}_L[\gamma_\ell^2]) = \log \mathcal{L}(\mathbf{c}, \mathbf{d} | \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \mathbf{m}_L[\gamma_\ell^2]) + \log \mathcal{L}(\boldsymbol{\nu}, \boldsymbol{\theta} | \boldsymbol{\zeta}, \mathbf{m}_L[\gamma_\ell^2]) + \log \mathcal{L}(\boldsymbol{\zeta} | \mathbf{m}_L[\gamma_\ell^2]).$$

For the first term we use a BIC-like approximation such that:

$$-2 \log \mathcal{L}(\mathbf{c}, \mathbf{d} | \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \mathbf{m}_L[\gamma_\ell^2]) \simeq NM \log \mathrm{RSS}(\mathbf{c}, \mathbf{d} | \boldsymbol{\nu}, \boldsymbol{\theta}) + (ML + 1) \times \log(N),$$

with $\mathrm{RSS}(\mathbf{c}, \mathbf{d} | \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta})$ the Residual Sum of Squares defined such that:

$$\mathrm{RSS}(\mathbf{c}, \mathbf{d} | \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\zeta}) = \sum_{i\ell} \zeta_{i\ell} \left\| \mathbf{c}_i - \widehat{\boldsymbol{\alpha}}_\ell - \boldsymbol{\nu}_{i\ell} \right\|^2 + \sum_{i\ell} \zeta_{i\ell} \left\| \mathbf{d}_i - \widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\theta}_{i\ell} \right\|^2.$$

Then we derive the integrated log-likelihood of the random effects. We assume a non-informative Jeffrey prior for the variance parameters such that $g(\gamma_{\nu,\ell}^2 | \boldsymbol{\zeta}, \mathbf{m}_L[\gamma_\ell^2]) \propto 1/\gamma_{\nu,\ell}^2$.

Using notations $N_\ell = \sum_{i=1}^N \zeta_{i\ell}$ and $\mathrm{RSS}_\ell(\boldsymbol{\nu}, \boldsymbol{\zeta}) = \sum_{i=1}^N \zeta_{i\ell}\nu_{i,\ell}^2$, we get:

$$-2\log \mathcal{L}(\boldsymbol{\nu}|\boldsymbol{\zeta}, \mathbf{m}_L[\gamma_\ell^2]) \simeq \sum_\ell N_\ell \log \mathrm{RSS}_\ell(\boldsymbol{\nu}, \boldsymbol{\zeta}) - 2\sum_\ell \log\Gamma\left(\frac{N_\ell}{2}\right).$$

Similarly for the detail coefficients we get:

$$-2\log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\zeta}, \mathbf{m}_L[\gamma_\ell^2]) \simeq (M-1)\sum_\ell N_\ell \log \mathrm{RSS}_\ell(\boldsymbol{\theta}, \boldsymbol{\zeta}) - 2\sum_\ell \log\Gamma\left(\frac{N_\ell(M-1)}{2}\right).$$

Finally for the classification term a Dirichlet *prior* is assumed for $g(\boldsymbol{\pi}|\mathbf{m}_L)$ and the corresponding integrated likelihood is approximated such as

$$\log \mathcal{L}(\boldsymbol{\zeta}|\mathbf{m}_L[\gamma_\ell^2]) \simeq \sum_{\ell=1}^L N_\ell \log(N_\ell/N) - \frac{(L-1)}{2}\log(N).$$

The last step of this derivation is to replace hidden variables by their predictions provided by the EM algorithm. Random effects $(\boldsymbol{\nu}, \boldsymbol{\theta})$ are replaced by their BLUP $(\widehat{\boldsymbol{\nu}}, \widehat{\boldsymbol{\theta}})$, and label variables $\boldsymbol{\zeta}$ are replaced by their conditional expectation $\boldsymbol{\tau}$. Put together we obtain the following integrated classification likelihood criterion (ICL):

$$\begin{aligned}
-\frac{2}{N} \times \mathrm{ICL}(\mathbf{m}_L[\gamma_\ell^2]) &= M \log \mathrm{RSS}(\mathbf{c}, \mathbf{d}|\widehat{\boldsymbol{\nu}}, \widehat{\boldsymbol{\theta}}, \boldsymbol{\tau}) \\
&+ \sum_\ell \widehat{\pi}_\ell \left(\log \mathrm{RSS}_\ell(\widehat{\boldsymbol{\nu}}, \boldsymbol{\tau}) + (M-1)\log \mathrm{RSS}_\ell(\widehat{\boldsymbol{\theta}}, \boldsymbol{\tau})\right) \\
&- \frac{2}{N}\sum_\ell \left\{\log\Gamma\left(\frac{\widehat{N}_\ell}{2}\right) + \log\Gamma\left(\frac{\widehat{N}_\ell(M-1)}{2}\right)\right\} \\
&- 2\sum_{\ell=1}^L \widehat{\pi}_\ell \log(\widehat{\pi}_\ell) + \frac{(M+1)L}{N} \times \log(N).
\end{aligned}$$

4. Simulations and Comparison of methods

4.1 Definition of a general simulation framework

In this Section we propose to define a unified framework for synthetic data generation for functional mixed models and functional clustering models. Using this unified strategy different methods can be fairly compared based on appropriately simulated data. First we properly define the Signal to Noise Ratio (SNR) in the functional domain. The SNR is defined as the ratio of signal power to the power of the measurement noise corrupting the signal. In our case, the power of the signal is defined such as:

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{T}\int_{\frac{T}{2}}^{-\frac{T}{2}} \sum_\ell \pi_\ell \mathbb{E}\left[|\mu_\ell(t) + U_i(t)|\right]^2 \mathrm{d}t &= \frac{1}{M}\sum_{\ell=1}^L \pi_\ell \left(\sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k\ell}^2 + \sum_{j\geq j_0}\sum_{k=0}^{2^j-1} \beta_{jk\ell}^2\right) \\
&+ 2^{j_0}\gamma_\nu^2 + \frac{2^{j_0(1-\eta)}\gamma_\theta^2}{1-2^{(1-\eta)}}.
\end{aligned}$$

The derivation of such formula is given in the Appendix. Hence we need to control two terms: SNR_μ that accounts for the power of the fixed effects:

$$\mathrm{SNR}_\mu^2 = \frac{1}{M\sigma_E^2}\sum_{\ell=1}^{L}\pi_\ell\left(\sum_{k=0}^{2^{j_0}-1}\alpha_{j_0k\ell}^2+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}\beta_{jk\ell}^2\right),$$

and the power of the random effect. For this purpose we introduce parameter

$$\lambda_{\mathrm{U}} = \sigma_E^2/\left(\gamma_\nu^2+\frac{\gamma_\theta^2}{1-2^{(1-\eta)}}\right),$$

using an analogy with the λ parameter used in the EM algorithm. When performing simulations, SNR_μ usually lies in $\{0.1,1,3,5,7\}$ and λ_{U} varies in $\{1/4,1,4\}$ such that small values of λ_{U} indicate an important variance for the random effects. In practice we also choose $\gamma_\nu^2=\gamma_\theta^2$.

To build fixed effects for simulations we generalize the approach described in Amato and Sapatinas (2005) which uses the well-known synthetic functions `Blocks`, `Bumps`, `Heavisine` and `Doppler` originally proposed by Donoho and Johnstone (1994). We choose L fixed effects for each synthetic function classes with the following expressions for $t\in[0,1]$ and $\ell=1,\ldots,L$

$$\begin{aligned}
\mu_\ell^{\texttt{Blocks}}(t) &= 10\sum_{r=1}^{11}\left(1+\frac{1}{2}h_r^\ell\mathrm{sgn}(t-v_r^\ell)\right),\\
\mu_\ell^{\texttt{Bumps}}(t) &= \sum_{r=1}^{11}h_r^\ell/\left(1+\frac{|t-v_r|}{w_r^\ell}\right)^4,\\
\mu_\ell^{\texttt{Heavisine}}(t) &= 4\sin(4\pi t)-\mathrm{sgn}(t-v_1^\ell)-\mathrm{sgn}(v_2^\ell-t),\\
\mu_\ell^{\texttt{Doppler}}(t) &= \sqrt{t(1-t)}\sin\left(2.1\pi/(t-t_0^\ell)\right),
\end{aligned}$$

where for Blocks and Bumps v_r^ℓ are the locations of the jumps chosen ramdomly in $[0,1]$, h_r^ℓ are the heights of the jumps and w_r^ℓ are the width of the bumps. As for Heavisine, v_1^ℓ, v_2^ℓ stand for the locations of the two discontinuities and for Doppler t_0^ℓ is the phase randomly chosen in $[0,1]$.

Once parameters $(\mathrm{SNR}_\mu,\lambda_{\mathrm{U}},\{\mu_\ell(t)\}_\ell)$ have been chosen (*ie* values for σ_E^2,γ^2 and $\boldsymbol{\alpha}_\ell,\boldsymbol{\beta}_\ell$ are deduced), our simulation procedure is performed in the wavelet domain such that realizations of centered Gaussian distribution with variance $2^{-j\eta}\gamma^2$ are added to the fixed effect coefficients to account for inter-individual variability. Then Gaussian noise with variance $\sigma_\varepsilon^2=\sigma_E^2/M$ is added to account for measurement errors. This unified method ensures that both fixed and random effects lie in the same Besov space, as mentionned earlier, and observed signals $\mathbf{Y}_i(\mathbf{t})$ can be recovered using the inverse DWT. An example of such simulated data is given in Figure 1.

4.2 Simulation Design and indicators of performance

Since too many configurations could be explored using simulations, we propose to fix the number of individuals at $n = 50$, the number of groups at $L = 2$, the length of the signals at $M = 512$, and parameter η is set to 2. Then the simulation design explores the following configurations: $\text{SNR}_\mu \in \{0.1, 1, 3, 5, 7\}$, $\lambda_{\text{U}} \in \{1/4, 1, 4\}$, $\pi \in \{0.1, 0.25, 0.5\}$, each simulation being repeated 50 times. In terms of methods, we compete functional clustering models with or without mixed effects (FCMM/FCM, Functional Clustering Mixed Model/Functional Clustering Model), and we consider (or not) the dimension reduction method based on the union of coefficients. We compare these 4 methods to the functional clustering mixed model based on splines as proposed by James and Sugar (2003) whose `R` code is available on the web page of the authors [1]. Our purpose is to highlight the benefit of using wavelets when dealing with high dimensional data.

The performance of the clustering procedures are compared using the Empirical Error Rate (EER) defined by

$$EER = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell}^{L} \mathbb{I}\{\widehat{\zeta}_{i\ell} \neq \zeta_{i\ell}\},$$

where $\widehat{\zeta}_{i\ell}$ is the predicted class for individual i and $\zeta_{i\ell}$ is the true class. This criteria ranges from 0, for which no classification error is made to 1 which means that all individuals are misclassified. We finally consider the speed of execution of each procedure.

4.3 Clustering performance

Figure 2 presents the variations of the Empirical Error Rates according to SNR_μ and to the strength of the random effect (a small λ_{U} indicates a strong random effect). A general comment is that the Functional Clustering Mixed Model (FCMM) outperforms all methods in terms of EER compared with the Functional Clustering Model (FCM) and Splines. FCMM has two main advantages. First the modeling of functional random effects leads to a better identification of the informative structures in terms of clustering. Table 1 clearly shows that FCMM is the best method to estimate the variance of the residuals contrary to FCM that provides over-estimates (which leads to poor clustering performance).

Then dimension reduction increases the performance of FCMM by removing coefficients that are not informative with respect to clustering. This is not true for the Functional Clustering Model (FCM) for which dimension reduction increases the EER. This trend can be explained by the bad estimation of the error's variance when random effects are not considered in the model. The selection of the coefficients that all survived thresholding leads to worst estimators in the case of FCM but the impact is moderate on the FCMM (Table 1).

Our last point concerns the time of execution of each method. When dealing with high dimensional data, it is crucial to propose methods that show reasonnable computa-

1. `http://www-bcf.usc.edu/~gareth/`

tional time. Table 1 clearly shows that using wavelet-based Functional Clustering Models gives the best execution times, and even when random effects are considered, time of execution remains moderate (less than 10 minutes for $n = 50$ individuals and $M = 512$ positions). Splines are known to be poorly efficient in terms of computational efficiency. This issue becomes critical when dealing with functional models with many individuals. The size of our simulated datasets was the upper limit that could be analyzed by Splines, in particular due to memory constraints. To this extent, our `R` package `curvclust` is the only freely available software that performs curve clustering with functional random effects within a reduced amount of time in high dimension.

5. Applications

5.1 Mass Spectrometry data

We first consider a SELDI-TOF mass spectrometry dataset issued from a study on ovarian cancer (Petricoin et al., 2002). The sample set includes serum profiles of 162 subjects with ovarian cancer and 91 non-cancer control subjects. Each serum profile consists of 15154 recorded intensities corresponding to distinct m/z values. This data set was produced by the Ciphergen WCX2 protein chip. It is available through the Clinical Proteomics Programs Databank (`http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`, ovarian dataset 8-7-02). Before clustering, raw data are background corrected using a quantile regression procedure, and spectra are aligned using a procedure based on wavelets zero crossings (Antoniadis et al., 2007a). Then the ovarian cancer dataset is made of 8192 intensities within the range of m/z ratio [1500,14000], ratios below 1500 being discarded due to the effects of matrix. We compete wavelet-based functional clustering models on these data considering different random effect structures. Procedures are applied in a non supervised framework to retrieve the known labels (cancer/control) and comparisons are based on empirical error rate estimates (EER, Table 2). Note that the spline-based procedure of James and Sugar (2003) could not be applied on these data because of their too high dimensionality.

The first result is that empirical error rates are high for all methods and that the introduction of random effects slightly decreases the EER whatever the random effect structure (from 38% to $\sim 25\%$). To investigate the origins of such modest performance, we also performed clustering based on group-wise aligned spectra instead of global alignment (which should be done in the unsupervised context). Results are striking: when spectra are aligned according to known labels model $\mathbf{m}_2[\gamma^2_{jk}]$ results in one mismatch only (EER=0.4%). This results leads to the following conclusions. First spectra alignment is a challenge when performing subgroup discovery, and the task is much more difficult compared with supervised clustering for which labels are known. Indeed inaccuracy in spectra alignment could lead to artificial differences in individual serum profiles which decreases the performance of clustering. A promising (but challenging) perspective would be to perform clustering and alignment simultaneously. Moreover as wavelets have been shown to perform best for peak-detection/alignment (Yang et al., 2009), our

wavelet-based procedure for clustering would be a good starting point to integrate both strategies.

Then a second result is that best clustering performance are provided by a functional clustering mixed model for which the random effect has a covariance structure that depends on both scale and location (γ^2_{jk}). This implies that inter-individual variations occur at specific ranges of m/z values, which reinforces the importance of correct spectra alignment. Interestingly, only an important proportion of variance terms are close to zeros which would make the BLUPs sparse if dimension reduction was performed on random effects. Unfortunately, the task is difficult in the non-supervised setting since BLUPs can not be computed without the knowledge of group-specific means (which would be possible in the supervised setting). Thus dimension reduction for clustering using mixed functional model remains challenging and still needs to be investigated.

5.2 Comparative Genomic Hybridization data

In this last application we consider the clustering of breast-cancer tumors based on their copy number aberration profiles measured by array-based Comparative Genomic Hydridization (Fridlyand et al., 2006). Array CGH is a widely used technology that enables the characterization of genome-wide chromosomal aberrations using the microarray technology. Many statistical methods have been developed to analyze these data (van de Wiel et al., 2011). They are mainly based on segmentation methods to retrieve segments of homogeneous copy number along the genome.

Clustering individuals based on their CGH profiles is a very challenging issue and has already been considered to identify new subtypes of tumors (Chin et al., 2007). For now, subgroup discovery is mainly performed using hierarchical clustering based on segmentation results (Van Wieringen et al., 2008). However the inter-individual variability has never been quantified in these data, contrary to mass spectrometry for instance. Thus using our method for clustering with the Haar basis (piece-wise constant basis) is a way to perform subgroup discovery by considering random effects. In the Fridlyand et al. (2006) paper, the authors identified 3 main subtypes of breast cancer that differ with respect to level of genomic instability. Interestingly, Van Wieringen et al. (2008) re-analyzed the data and do not mention much correspondance between the two clustering results. Moreover, they discovered much more subgroups and noticed that "the samples in the study could be more heterogeneous than previously implied".

We also find more subgroups than the original study, with 5 clusters selected by ICL (2 by the BIC). First, this shows the power which is gained when considering the random effect in the selection step. Then we were able to identify the 1q/16p subtype on the complete dataset (with 1 mismatch). This subtype was identified in the first study (Fridlyand et al., 2006) but not by other clustering methods (Van Wieringen et al., 2008) whereas it is associated to the best patient outcome. Since 2 of the 3 identified clusters in the original paper concern ER positive tumors, we also performed our method on this subset of patients and retrieve the 1q/16p subtype without mismatch. In this classification, one cluster was made of 3 tumors (S0041, S0041, S1519) also identified as similar in the original paper. As a last resul Table 3 indicates that the

estimated signal to noise ratio is low and the impressive strength of the random effect ($\widehat{\lambda}_{\mathrm{U}} \sim 10^{-4}$) also indicates that the inter-individual variability is ultra-high in these data. As a consequence, finding clusters with biological significance will require rather hundreds/thousands of patients compared with 55 in the original study.

6. Conclusion

In this work we provide a methodology for model-based clustering of functional data in the presence of inter-individual variability. Our method is based on a wavelet decomposition of the signal and on a mixture model that integrates random effects. We illustrate the power of such an approach in two different fields of high-throughput biology using our package `curvclust`, and we show the potentialities of functional models on array CGH data. Overall, random effects allow us to properly model the variance structure of the data, and to exhibit the high proportion of variance due to inter-individual variability. This part is usually omitted in high-throughput modelling. First perspective will concern the generalization of our approach to the supervised setting. Finding biomarkers has received enormous attention in the past years, with moderate success due to the lack of reproducibility. Our study in the non-supervised framework shows that the inter-individual variability is important in these data, which may be one explanation of the difficulty to find reliable markers. Integrating random effects in the supervised setting may produce more moderate results, but at least they would be more representative of the biological variability. Finally methodological perspectives of this work will mainly concern dimension reduction. The task is difficult in the non-supervised setting and the illustration on MS data shows that dimension reduction should be performed for fixed *and* for random effects which remains challenging. This would provide a better representation of the signal by thresholding coefficients with poor information, and would increase the speed of the estimation algorithm that is sensitive to the number of selected coefficients, which is of central interest of high dimensional data.

Acknowledgements

Part of this work was supported by the Interuniversity Attraction Pole (IAP) research network in Statistics P5/24.

References

F. Abramovich, T. Sapatinas, and B.W. Silverman. Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society Series B Stat Methodol*, 60:725–749, 1998.

U. Amato and T. Sapatinas. Wavelet shrinkage approaches to baseline signal estimation from repeated noisy measurements. *Advances and Applications in Statistics*, 51:21–50, 2005.

A. Antoniadis and T. Sapatinas. Estimation and inference in functional mixed-effects models. *Computational Statistics & Data Analysis*, 51(10):4793–4813, 2007b.

A. Antoniadis, J. Bigot, S. Lambert-Lacroix, and F. Letue. Non parametric preprocessing methods and inference tools for analyzing time-of-flight mass spectrometry data. *Current Analytical Chemistry*, 3(2):127–147, 2007a.

A. Antoniadis, J. Bigot, and R. von Sachs. A multiscale approach for statistical characterization of functional images. *Journal of Computational and Graphical Statistics*, 18(1):216–237, 2008.

H. Bensmail, B. Aruna, O. J. Semmes, and A. Haoudi. Functional clustering algorithm for high-dimensional proteomics data. *J. Biomed. Biotechnol.*, 2005:80–86, Jun 2005.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE PAMI*, 22(7):719–725, 2000.

S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M. A. van de Wiel, A. R. Green, I. O. Ellis, P. L. Porter, S. Tavare, J. D. Brenton, B. Ylstra, and C. Caldas. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, 8:R215, 2007.

D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

D.L. Donoho and I.M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26:879–921, 1998.

J. E. Eckel-Passow, A. L. Oberg, T. M. Therneau, and H. R. Bergen. An insight into high-resolution mass-spectrometry data. *Biostatistics*, 10:481–500, Jul 2009.

J. Fan. Test of significance based on wavelet thresholding and Neymans truncation. *JASA*, 91:674–688, 1996.

J. Fridlyand, A. M. Snijders, B. Ylstra, H. Li, A. Olshen, R. Segraves, S. Dairkee, T. Tokuyasu, B. M. Ljung, A. N. Jain, J. McLennan, J. Ziegler, K. Chin, S. Devries, H. Feiler, J. W. Gray, F. Waldman, D. Pinkel, and D. G. Albertson. Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, 6:96, 2006.

M. Hilario, A. Kalousis, C. Pellegrini, and M. Muller. Processing and classification of protein mass spectra. *Mass Spectrom Rev*, 25:409–449, 2006.

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

J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953.

M. Kurodaa and M. Sakakiharab. Accelerating the convergence of the EM algorithm using the vector epsilon algorithm. *Computational Statistics & Data Analysis*, 51: 1549–1561, 2006.

J. S. Morris and R. J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B Stat Methodol*, 68:179–199, 2006.

J. S. Morris, P. J. Brown, R. C. Herrick, K. A. Baggerly, and K. R. Coombes. Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64:479–489, Jun 2008.

J. S. Morris, K. A. Baggerly, H. B. Gutstein, and K. R. Coombes. Statistical contributions to proteomic research. *Methods Mol. Biol.*, 641:143–166, 2010.

E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359:572–577, Feb 2002.

S. Ray and B. Mallick. Functional clustering by bayesian wavelet methods. *Journal of the Royal Statistical Society Series B Stat Methodol*, 68(2):305–332, 2006.

M. A. van de Wiel, F. Picard, W. N. van Wieringen, and B. Ylstra. Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief. Bioinformatics*, 12:10–21, Jan 2011.

W. N. Van Wieringen, M. A. Van De Wiel, and B. Ylstra. Weighted clustering of called array CGH data. *Biostatistics*, 9:484–500, Jul 2008.

C. Yang, Z. He, and W. Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics*, 10(4):1–13, 2009.

7. Appendix

7.1 Derivation of the signal power

Considering compactly supported functions on $[0,1]$ and a centered Gaussian process for $U_i(t)$, the mean power of the signal is derived such that:

$$\lim_{T\to\infty}\frac{1}{T}\int_{\frac{T}{2}}^{-\frac{T}{2}}\sum_{\ell}\pi_\ell\mathbb{E}\big[|\mu_\ell(t)+U_i(t)|\big]^2\mathrm{d}t \;=\; \sum_{\ell}\pi_\ell\int_0^1|\mu_\ell(t)|^2\mathrm{d}t+\sum_{\ell}\pi_\ell\int_0^1\mathbb{E}\big[U_i(t)^2\big]\mathrm{d}t.$$

Using the law of the conservation of energy, the power of the fixed effects is:

$$\sum_{\ell}\pi_\ell\int_0^1|\mu_\ell(t)|^2\mathrm{d}t=\frac{1}{M}\sum_{\ell=1}^{L}\pi_\ell\left(\sum_{k=0}^{2^{j_0}-1}\alpha_{j_0k\ell}^2+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}\beta_{jk\ell}^2\right).$$

As for the power of the random effects it is derived using the orthonormality of wavelet basis

$$\begin{aligned}\int_0^1\mathbb{E}[U_i(t)^2]\mathrm{d}t&=\int_0^1\mathbb{E}\left[\left(\sum_{k=0}^{2^{j_0}-1}\nu_{ij_0k}\phi_{j_0k}(t)+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}\theta_{ijk}\psi_{jk}(t)\right)^2\right]\mathrm{d}t\\&=\sum_{k=0}^{2^{j_0}-1}\gamma_\nu^2+\sum_{j\geq j_0}\sum_{k=0}^{2^j-1}2^{-j\eta}\gamma_\theta^2\\&=2^{j_0}\gamma_\nu^2+\frac{2^{j_0(1-\eta)}\gamma_\theta^2}{1-2^{(1-\eta)}}.\end{aligned}$$

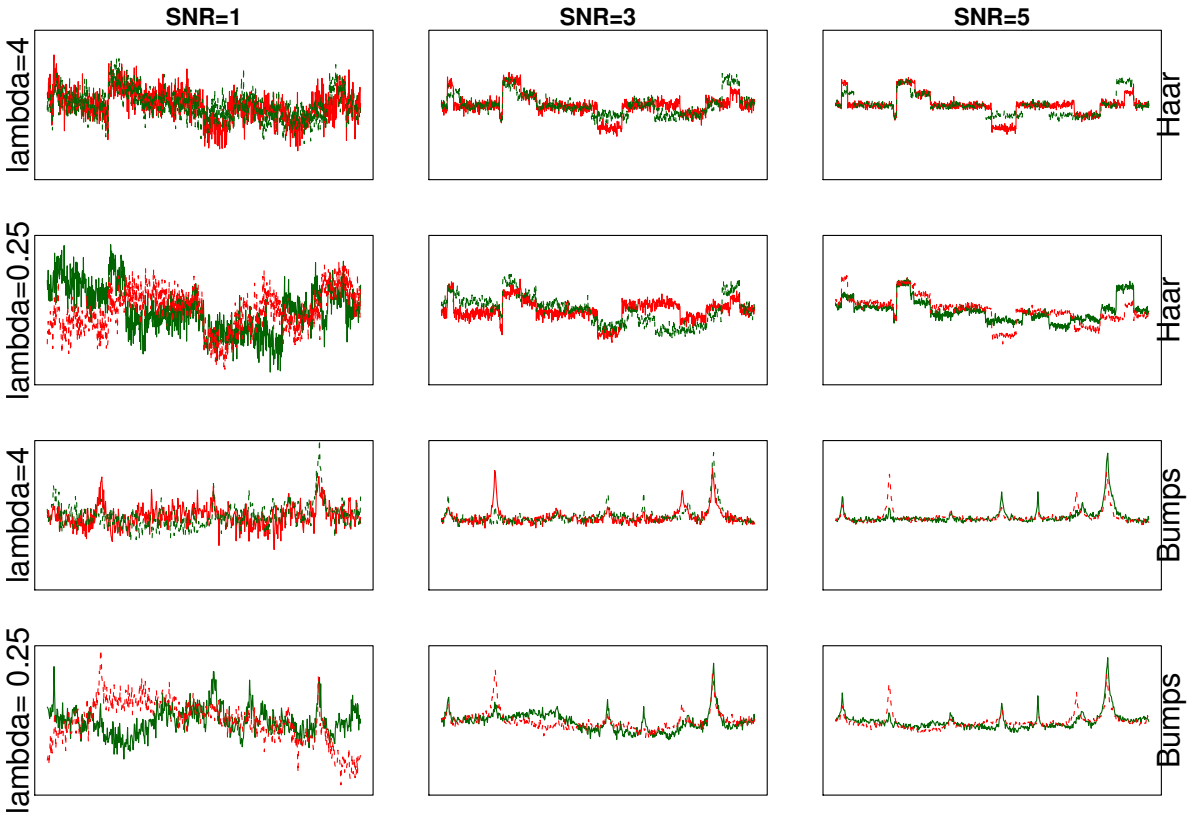

Figure 1: Example of simulated curves with varying SNR_μ and λ_U (One curve per cluster).

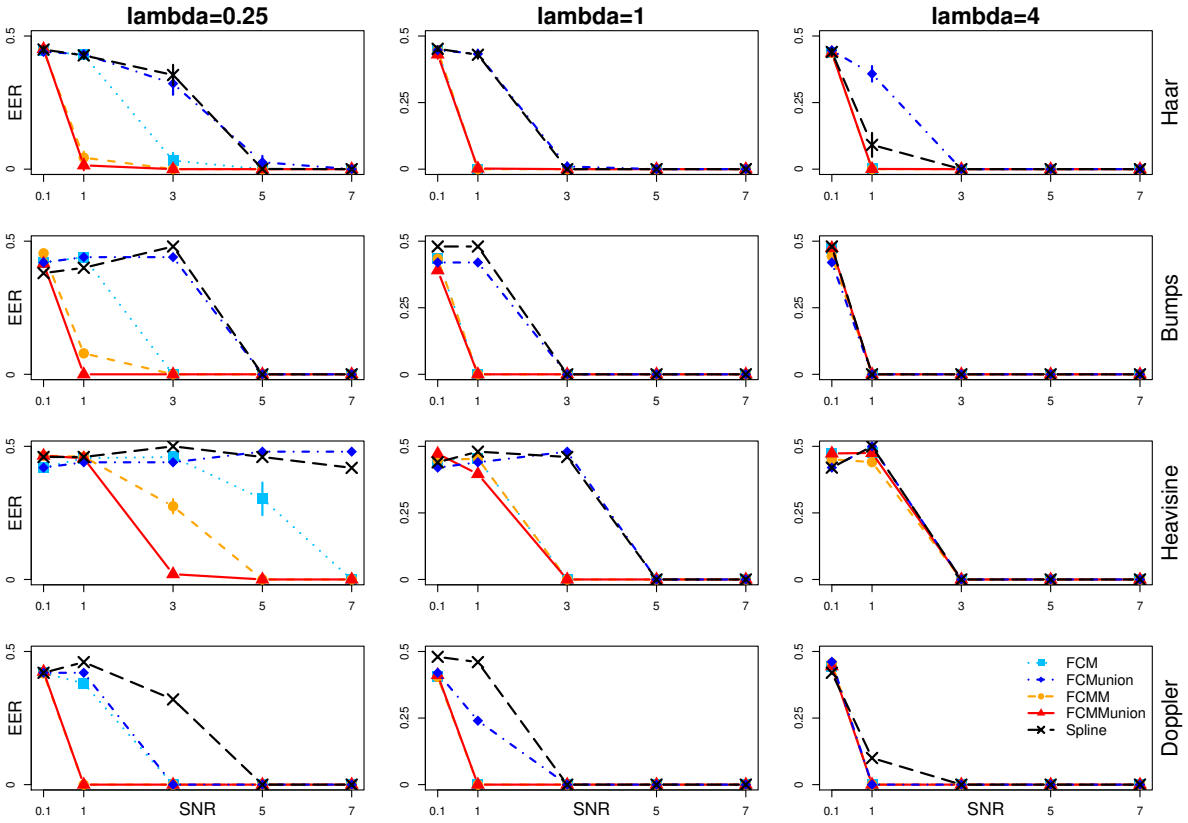

Figure 2: Variation of the Empirical Error Rate (EER) for different estimation methods: Functional Clustering Mixed Model (FCMM), Functional Clustering Model (FCM), with or without dimension reduction ('union'), and Splines. In columns different intensities for the variance of the random effect are considered: $\lambda_U = 0.25/1/4$ for a strong/mild/small random effect. In rows are considered different shapes for the mean curve of each group (Haar, Bumps, Heavisine, Doppler).

		Bias					TOE				
SNR^2_μ		0.1	1	3	5	7	0.1	1	3	5	7
FCM	Haar	-2.57	-2.66	-2.96	-3.02	-2.99	2.3	2.4	2.3	2.4	2.3
	Bumps	-2.50	-2.69	-2.93	-2.93	-2.93	2.6	2.5	2.6	2.5	2.5
	Heavisine	-2.15	-2.17	-3.22	-4.30	-2.50	2.8	2.7	2.7	2.7	2.8
	Doppler	-2.73	-3.07	-3.32	-3.33	-3.33	2.9	3.2	3.1	3.2	3.2
FCMu	Haar	-12.93	-11.33	-9.42	-9.38	-8.89	0.4	0.4	0.5	0.5	0.5
	Bumps	-12.98	-11.11	-13.46	-11.98	-11.93	0.5	0.5	0.5	0.5	0.5
	Heavisine	-11.62	-10.20	-10.07	-12.05	-15.68	0.5	0.5	0.5	0.5	0.5
	Doppler	-14.75	-13.14	-11.33	-8.59	-7.87	0.5	0.5	0.5	0.6	0.6
FCMM	Haar	0.11	0.05	-0.01	-0.01	-0.00	16.0	16.1	15.6	15.8	16.0
	Bumps	0.09	0.04	0.01	0.01	0.01	16.1	16.3	15.2	15.3	15.4
	Heavisine	0.10	0.09	0.08	0.03	0.02	16.4	16.2	16.0	16.4	15.9
	Doppler	0.08	0.01	-0.02	-0.02	-0.01	17.5	17.4	17.5	16.4	17.0
FCMMu	Haar	-0.11	-0.06	0.03	0.06	0.05	6.9	7.1	7.6	7.6	7.6
	Bumps	-0.10	-0.04	-0.08	-0.08	-0.05	6.7	6.7	6.8	6.7	6.7
	Heavisine	-0.10	-0.10	-0.18	-0.21	-0.19	7.1	7.3	6.8	6.8	6.8
	Doppler	-0.18	-0.06	-0.04	-0.16	-0.11	7.3	7.1	7.3	7.8	7.9
Spline	Haar	.	.	.	.	.	25.5	26.2	23.0	23.6	22.3
	Bumps	.	.	.	.	.	23.3	26.6	22.0	21.2	21.7
	Heavisine	.	.	.	.	.	24.2	21.6	21.8	22.4	22.3
	Doppler	.	.	.	.	.	33.2	32.4	24.2	24.8	24.2

Table 1: Relative bias of the estimator of the error variance: $(\sigma^2 - \widehat{\sigma}^2)/\sigma^2$, and average time of execution in minutes for different models on simulated data ($n = 50$ individuals, $M = 512$ positions). FCM, functional clustering model, FCMM functional clustering mixed model. FCMu/FCMMu: functional clustering (mixed) models based on the union of coefficients for dimension reduction. Programs were run on a cluster of 2 octo-bicore Opteron 2.8Ghz and 2 octo-quadcore Opteron 2.3GHz.

	$\mathbf{m}_2$	$\mathbf{m}_2[\gamma^2]$	$\mathbf{m}_2[\gamma^2_\ell]$	$\mathbf{m}_2[\gamma^2_{jk}]$	$\mathbf{m}_2[\gamma^2_{jk\ell}]$
global alignment	38	24	24	23	23
group alignment	20	21	22	0.4	36

Table 2: Empirical Error Rates (in percent) for the Petricoin et al. (2002) data for different models: functional clustering without random effects, 2 groups ($\mathbf{m}_2$), functional clustering with random effect with different variance structures for the random effect: constant $\mathbf{m}_2[\gamma^2]$, group $\mathbf{m}_2[\gamma^2_\ell]$, scale-position $\mathbf{m}_2[\gamma^2_{jk}]$, or group-scale-position dependent $\mathbf{m}_2[\gamma^2_{jk\ell}]$.

	Complete dataset			ER+ dataset	
cluster ID	$\widehat{\mathrm{SNR}}_\mu^2$	$\widehat{\lambda}_\mathrm{U}$	cluster ID	$\widehat{\mathrm{SNR}}_\mu^2$	$\widehat{\lambda}_\mathrm{U}$
1	2.1e-4	3.9e-04	1	2.1e-3	2.2e-04
2	2.3e-3	3.8e-05	2	7.8e-3	1.9e-05
3	1.3e-3	6.4e-04	3	1.1e-2	3.8e-05
4 (1q/16p)	1.5e-3	1.3e-04	4 (1q/16p)	4.4e-3	4.4e-04
5	9.3e-4	4.3e-05			

Table 3: Estimated SNR_μ^2 and λ_U for the breast tumor dataset of Fridlyand et al. (2006).