The Discriminative Functional Mixture Model for the Analysis of Bike Sharing Systems

Charles Bouveyron, Etienne Côme, Julien Jacques

Submitted to the Annals of Applied Statistics

THE DISCRIMINATIVE FUNCTIONAL MIXTURE MODEL FOR THE ANALYSIS OF BIKE SHARING SYSTEMS

Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes[*],*
Laboratoire GRETTIA, IFSTTAR[†],
Laboratoire Paul Painlevé, UMR CNRS 8524, Université Lille 1 & INRIA[‡]

By Charles Bouveyron[*], Etienne Côme[†], and Julien Jacques[‡]

Bike sharing systems (BSSs) have become a mean of sustainable intermodal transport and are now proposed in many cities worldwide. Most of BSSs also provide an open access to their data, through APIs, and in particular an access to real time status reports on their bike stations. The analysis of the data generated by such systems is of particular interest for BSS providers, researchers in social sciences and even users. This work was motivated by the analysis and the comparison of several European BSSs. To this end, a model-based clustering method, called FunFEM, for time series (or more generally functional data) is developed. It is based on a discriminative functional mixture model which allows the clustering of the data in a functional subspace. This model presents the advantage to be parsimonious and can therefore handle long time series. Numerical experiments confirm the good behavior of FunFEM, in particular compared to state-of-the-art methods. The application of FunFEM to BSS data from JCDecaux and Transport for London Initiative provides insightful analyses on the Vélib system and informative comparisons between 8 European BSSs.

1. Introduction. The recent emergence of open data systems and their interface with personal devices through APIs (application programming interfaces) allows people to access to useful information as wide as weather data and forecasts, university course schedules, real-time bus time tables or bike availabilities in bike sharing systems (BSSs). Beyond these useful and practical aspects of open data, it is of great interest for system providers and scientists to analyze these data in order to understand how people use the studied system and help in improving the user experience. This work was motivated by the analysis of bike sharing systems which provide real time status reports on their bike stations (number of available bikes, number of free bike stands, ...) for dozens of cities worldwide.

The implementation of bike sharing systems is one of the urban mobility services proposed in many cities all over the world as an additional mean of sustainable intermodal transport. Over the last few years, different BSSs have been implemented in European, American and Chinese cities. The main

motivation is to provide users with free or rental bicycles especially suited for short-distance trips in urban areas, thus reducing traffic congestion, air pollution and noise. Thanks to their unquestionable success (De Maio, 2009; Büttner et al., 2011), more and more cities want to provide this type of mobility service in order to show that they are sustainable and modern. In France, since the implementation of the first BSS in Lyon in 2005 (the Vélo'v system), BSSs have been launched in twenty French cities, including Paris, one of the most large-scale BSS implemented in France (the Vélib' system).

The key of the success is to have a good knowledge of BSS usage and performance. This knowledge can then be transferred to other cities wishing to introduce BSSs. Several studies (Froehlich, Neumann and Oliver, 2009; Borgnat et al., 2011; Vogel and Mattfeld, 2011; Lathia, Saniul and Capra, 2012) have shown the usefulness of analyzing the data collected by BSS operators and city authorities. A statistical analysis of such data helps in the development of new and innovative approaches for a better understanding of both urban mobility and BSS use. The design of BSSs, the adjustment of pricing policies, the improvement of service level of the system (redistribution of bikes over stations) can all benefit from this kind of analyses (Dell'Olio, Ibeas and Moura, 2011; Lin and Yang, 2011; APUR, 2006). Additionally, such analyses help sociologists and BSS providers to understand user mobility patterns within the cities.

However, the amount of data collected on such systems is often very large. It is therefore difficult to acquire knowledge using it without the help of automatic algorithms which extract mobility patterns and give a synthetic view of the information. The clustering of the BSS stations is closely related to the city activities (transportation, leisure, employment) and can be helpful for a variety of applications, including urban planning and business localization. In addition, the analysis of the results provides insights into the relations between the neighborhood types and their associated usage profiles.

In almost all of the clustering studies carried out until now, the bicycle sharing stations are grouped according to their usage profiles, thus highlighting the relationships between time of day, location and usage. The first attempt in this line of works is due to Froehlich, Neumann and Oliver (2008), who analyzed a data set from the Barcelona Bicing system. The data correspond to station occupancy statistics in the form of free slots, available bikes over several time frames and other station activity statistics derived from station occupancy data collected every 5 minutes. The clustering is performed using a Gaussian mixture model based on features such as the average number of available bikes at different periods of the day. It should be noticed that such techniques do not really take advantage of the temporal

dynamic of data. In Froehlich, Neumann and Oliver (2009), two clusterings are compared, both being performed by hierarchical aggregation. The first one uses activity statistics derived from the evolution of station occupancy while the second uses directly the number of available bicycles along the day. Other studies, such as Lathia, Saniul and Capra (2012), use similar clustering techniques and data to study the effect of changing the user-access policy in the London Barclays cycle hire scheme. The authors investigate how this change has impacted the system usage throughout the city via both spatial and temporal analysis of station occupancy data. As in Froehlich, Neumann and Oliver (2009), each station is described by a time series vector which corresponds to the normalized available bicycle value of the station along the day. Each element of the feature vector is therefore equal to the number of available bicycles divided by the station size. These time series are then smoothed using a moving average and clustered using a hierarchical agglomerative algorithm (Duda, Hart and Stork, 2001, see p. 552), with a cosine distance. Another work that uses the same type of data was proposed by Vogel and Mattfeld (2011); Vogel, Greiser and Mattfeld (2011); it aims at identifying station clusters in order to better understand temporal and spatial causes of imbalances between BSS stations. The proposed methodology, based on the Geographical Business Intelligence process, was successfully applied to data collected from Vienna's BSS (Citybike Wien). It uses feature vectors to describe the stations that come from normalizing arrival and departure counts per hour, and also handles weekdays and weekends separately. Classical clustering algorithms, namely k-means, Gaussian mixture model and sequential Information-Bottleneck (sIB), are then compared. Finally, Côme and Oukhellou (2014) recently proposed an original approach considering a generative model based on Poisson mixtures to cluster stations with respect to hourly usage profiles build from trip data. The results obtained for the Vélib' system (Paris) were then analyzed with respect to the city geography and sociology.

In all these works, the observed time series are clustered using either geometric methods based on distances between time series, or by creating features summarizing the activity in the given periods of the day (and thus omitting the temporal dynamics of the data). Recent advances in functional data analysis have contributed to develop efficient clustering techniques specific to functional data. One of the earlier works in that domain is due to James and Sugar (2003) who define an approach particularly effective for sparsely sampled functional data. This method, called fclust, considers that the basis expansion coefficients of the curves into a spline basis are distributed according to a mixture of Gaussian distributions with cluster-

specific means and common variances. The use of a spline basis is convenient when the curves are regular, but are not appropriate for peak-like data as encountered in mass spectrometry for instance. For this reason, Giacofci et al. (2012) recently proposed a Gaussian model on a wavelet decomposition of the curves. This approach allows to deal with a wider range of functional shapes than splines. An interesting approach has also been considered in Samé et al. (2011) who assume that the curves arise from a mixture of regressions on a basis of polynomial functions, with possible regime changes at each instant of observation. Let also mention the work of Frühwirth-Schnatter and Kaufmann (2008) who have built a specific clustering algorithm based on parametric time series models. Bouveyron and Jacques (2011) have extended the high-dimensional data clustering (HDDC) algorithm (Bouveyron, Girard and Schmid, 2007) to the functional case. The resulting model assumes a parsimonious cluster-specific Gaussian distribution for the basis expansion coefficients. More recently, Jacques and Preda (2013) have proposed a model-based clustering built on the approximation of the notion of density for functional variables, extended to multivariate functional data in Jacques and Preda (2014). These models assume that the functional principal component scores of curves have a Gaussian distribution whose parameters are cluster-specific. Some Bayesian approaches have also been proposed. On the one hand, Heard, Holmes and Stephens (2006) have considered that the basis expansion coefficients are distributed as a mixture of Gaussian whose variances are modeled by an Inverse-Gamma distribution. On the other hand, Ray and Mallick (2006) propose a nonparametric Bayes wavelet model for curve clustering based on a mixture of Dirichlet processes.

In this work, a novel model-based clustering method for time series (and more generally functional data) is proposed. This method, called FunFEM, is based on the discriminative functional mixture (DFM) model which models the data into a single discriminative functional subspace. This subspace allows afterward an insightful visualizations of the clustered data. A family of 12 models is also proposed by relaxing or constraining the main DFM model, allowing to handle a wide range of situations. The FunFEM algorithm is proposed for the inference of the DFM models, and model selection can be performed either by BIC or the "slope heuristic". Additionally, the selection of the most discriminative basis functions can be done afterward by introducing some sparsity through a ℓ_1-type penalization. The paper is organized as follows. Section 2 introduces the DFM model, its model family and the FunFEM algorithm. Model choice and selection of the discriminative functions are also discussed in Section 2. Numerical experiments on simu-

lated and benchmark data sets are then presented in Section 3 for validating the proposed approach. Section 4 presents the analyses and comparisons of 8 bike sharing systems using the FunFEM algorithm. Section 5 finally provides some concluding remarks.

2. The discriminative functional mixture model. In this work, we aim at clustering a set of observed curves $\{x_1, ..., x_n\}$ into K homogenous groups (or clusters) allowing the analysis of the studied process. To this end, this section introduces a latent functional model which adapts the model of Bouveyron and Brunet (2012) proposed in the multivariate case. An original inference algorithm for the functional model is then proposed, allowing afterward the clustering of the curves. Model choice and variable selection are also discussed.

2.1. *Transformation of the observed curves.* Let us first assume that the observed curves $\{x_1, ..., x_n\}$ are independent realizations of a L_2-continuous stochastic process $X = \{X(t)\}_{t\in[0,T]}$ for which the sample paths, *i.e.* the observed curves, belong to $L_2[0, T]$. In practice, the functional expressions of the observed curves are not known and we only have access to the discrete observations $x_{ij} = x_i(t_{is})$ at a finite set of ordered times $\{t_{is} : s = 1, \ldots, m_i\}$. It is therefore necessary to first reconstruct the functional form of the data from their discrete observations. A common way to do this is to assume that curves belong to a finite dimensional space spanned by a basis of functions (see for example Ramsay and Silverman, 2005). Let us therefore consider such a basis $\{\psi_1, \ldots, \psi_p\}$ and assume that the stochastic process X admits the following basis expansion:

$$X(t) = \sum_{j=1}^{p} \gamma_j(X)\psi_j(t), \tag{2.1}$$

where $\gamma = (\gamma_1(X), ..., \gamma_p(X))$ is a random vector in $\mathbb{R}^p$ and the number p of basis functions is assumed to be fixed and known. The basis expansion of each observed curve $x_i(t) = \sum_{j=1}^{p} \gamma_{ij}\psi_j(t)$ can be estimated by an interpolation procedure (see Escabias, Aguilera and Valderrama (2005) for instance) if the curves are observed without noise, or by least square smoothing if they are observed with error:

$$x_i^{obs}(t_{is}) = x_i(t_{is}) + \varepsilon_{is} \quad s = 1, \ldots, m_i.$$

The latter option is used in the present work. In this case, the basis coefficients of each sample path x_i are approximated by

$$\widehat{\gamma}_i = \left(\Theta_i'\Theta_i\right)^{-1}\Theta_i' X_i^{obs},$$

with $\Theta_i = (\psi_j(t_{is}))_{1\leq i\leq n, 1\leq s\leq m_i}$ and $X_i^{obs} = (x_i^{obs}(t_{i_1}), \ldots, x_i^{obs}(t_{i_{m_i}}))'$.

2.2. *The model.* The goal is to cluster the observed curves $\{x_1, ..., x_n\}$ into K homogeneous groups. Let us assume that there exists an unobserved random variable $Z = (Z_1, \ldots, Z_K) \in \{0, 1\}^K$ indicating the group membership of X: Z_k is equal to 1 if X belongs to the kth group and 0 otherwise. The clustering task aims therefore to predict the value $z_i = (z_{i1}, \ldots, z_{iK})$ of Z for each observed curve x_i, for $i = 1, ..., n$.

Let $F[0, T]$ be a latent subspace of $L_2[0, T]$ assumed to be the most discriminative subspace for the K groups spanned by a basis of d basis functions $\{\varphi_j\}_{j=1,...,d}$ in $L_2[0, T]$, with $d < K < p$. The basis $\{\varphi_j\}_{j=1,...,d}$ is obtained from $\{\psi_j\}_{j=1,...,p}$ through a linear transformation $\varphi_j = \sum_{\ell=1}^{p} u_{j\ell}\psi_\ell$ such that the $p \times d$ matrix $U = (u_{j\ell})$ is orthogonal. Let $\{\lambda_1, ..., \lambda_n\}$ be the latent expansion coefficients of the curves $\{x_1, ..., x_n\}$ in the basis $\{\varphi_j\}_{j=1,...,d}$. These coefficients are assumed to be independent realizations of a latent random vector $\Lambda \in \mathbb{R}^d$. The relationship between both bases $\{\varphi_j\}_{j=1,...,d}$ and $\{\psi_j\}_{j=1,...,p}$ suggests that the random vectors Γ and Λ are linked through the following linear transformation:

$$\Gamma = U\Lambda + \varepsilon, \tag{2.2}$$

where $\varepsilon \in \mathbb{R}^p$ is an independent and random noise term.

Let us now make some distributional assumptions on the random vectors Λ and ε. Firstly, conditionally on Z, Λ is assumed to be distributed according to a multivariate Gaussian density:

$$\Lambda_{|Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k), \tag{2.3}$$

where m_k and S_k are respectively the mean and the covariance matrix of the kth group. Secondly, ε is also assumed to be distributed according to a multivariate Gaussian density:

$$\varepsilon \sim \mathcal{N}(0, \Xi). \tag{2.4}$$

With these distributional assumptions, the marginal distribution of Γ is a mixture of Gaussians:

$$p(\gamma) = \sum_{k=1}^{K} \pi_k \phi(\gamma; U\mu_k, U^t\Sigma_k U + \Xi), \tag{2.5}$$

where ϕ is the standard Gaussian density function and $\pi_k = P(Z = k)$ is the prior probability of the kth group.

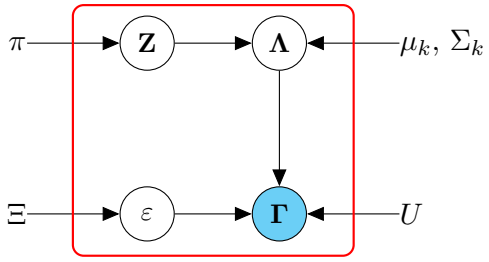

FIG 2.1. *Graphical representation for the model* $\mathrm{DFM}_{[\Sigma_k\beta]}$.

We finally assume that the noise covariance matrix Ξ is such that $\Delta_k = \mathrm{cov}(W^t\Gamma|Z=k) = W^t\Sigma_k W$ has the following form:

$$\Delta_k = \begin{pmatrix} \boxed{\Sigma_k} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} \beta & & 0 \\ & \ddots & \\ 0 & & \beta \end{matrix}} \end{pmatrix} \begin{matrix} \} \, d \\ \} \, p-d \end{matrix} \tag{2.6}$$

with $W = [U, V]$ where V is the orthogonal complement of U. With these notations and from a practical point of view, one can say that the variance of the actual data of the kth group is therefore modeled by Σ_k whereas the parameter β models the variance of the noise outside the functional subspace. This model is referred in the sequel by $\mathrm{DFM}_{[\Sigma_k\beta]}$ and Figure 2.1 summarizes the modeling.

2.3. *A family of discriminative functional model.* Starting with the model $\mathrm{DFM}_{[\Sigma_k\beta]}$ and following the strategy of Bouveyron and Brunet (2012), several submodels can be generated by applying constraints on parameters of the matrix Δ_k. For instance, it is first possible to relax the constraint that the noise variance is common across the groups. This generates the model $\mathrm{DFM}_{[\Sigma_k\beta_k]}$ which is the more general model of the family. It also possible to constrain this new model such that the covariance matrices $\Sigma_1, \ldots, \Sigma_K$ in the latent space are common across groups. This submodel will be referred by $\mathrm{DFM}_{[\Sigma\beta_k]}$. Similarly, in each group, Σ_k can be assumed to be diagonal, *i.e.* $\Sigma_k = \mathrm{diag}(\alpha_{k1}, \ldots, \alpha_{kd})$ and this submodel will be referred by $\mathrm{DFM}_{[\alpha_{kj}\beta_k]}$. The variance within the latent subspace F can also be assumed to be isotropic for each group and the associated submodel is $\mathrm{DFM}_{[\alpha_k\beta_k]}$. Following this strategy, 12 different DFM models can be enumerated and

Model	Σ_k	β_k	Nb. of variance parameters
$\text{DFM}_{[\Sigma_k \beta_k]}$	Free	Free	$(K-1)(p-K/2)+K^2(K-1)/2+K$
$\text{DFM}_{[\Sigma_k \beta]}$	Free	Common	$(K-1)(p-K/2)+K^2(K-1)/2+1$
$\text{DFM}_{[\Sigma \beta_k]}$	Common	Free	$(K-1)(p-K/2)+K(K-1)/2+K$
$\text{DFM}_{[\Sigma \beta]}$	Common	Common	$(K-1)(p-K/2)+K(K-1)/2+1$
$\text{DFM}_{[\alpha_{kj} \beta_k]}$	Diagonal	Free	$(K-1)(p-K/2)+K^2$
$\text{DFM}_{[\alpha_{kj} \beta]}$	Diagonal	Common	$(K-1)(p-K/2)+K(K-1)+1$
$\text{DFM}_{[\alpha_k \beta_k]}$	Spherical	Free	$(K-1)(K-1)(p-K/2)+2K$
$\text{DFM}_{[\alpha_k \beta]}$	Spherical	Common	$(K-1)(p-K/2)+K+1$
$\text{DFM}_{[\alpha_j \beta_k]}$	Diagonal & Common	Free	$(K-1)(p-K/2)+(K-1)+K$
$\text{DFM}_{[\alpha_j \beta]}$	Diagonal & Common	Common	$(K-1)(p-K/2)+(K-1)+1$
$\text{DFM}_{[\alpha \beta_k]}$	Spherical & Common	Free	$(K-1)(p-K/2)+K+1$
$\text{DFM}_{[\alpha \beta]}$	Spherical & Common	Common	$(K-1)(p-K/2)+2$

TABLE 1

Number of free parameters in covariance matrices when $d = K-1$ for the DFM models.

an overview of them is proposed in Table 1. The table also provides, for each model, the number of variance parameters to estimate as a function of the number K of groups and the number p of basis functions. One can notice that the models turn out to be particularly parsimonious since their complexity is a linear function of p whereas most model-based approaches usually have a complexity which is a quadratic function of p.

2.4. *Model inference: the FunFEM algorithm.* Since the group memberships $\{z_1, ..., z_n\}$ for the curves are unknown, the direct maximization of the likelihood associated with the model described above is intractable. In such a case, a classical solution for model inference is to use the EM algorithm. Here, however, the use of the EM algorithm is prohibited due to the particular nature of the functional subspace F. Indeed, maximizing the likelihood over the subspace orientation matrix U is equivalent to maximizing the projected variance and yields the functional principal component analysis (fPCA) subspace. Since F is here assumed to be the most discriminative subspace, U has to be estimated separately and we therefore propose the algorithm described hereafter and named FunFEM. The FunFEM algorithm alternates, at iteration q, over the three following steps:

The F step. Let us first suppose that at, iteration q, the posterior probabilities $t_{ik}^{(q)} = E[z_{ik}|\gamma_i, \theta^{(q-1)}]$ are known (they have been estimated in the E step of iteration $q-1$). The F step aims therefore at determining, conditionally to the $t_{ik}^{(q)}$, the orientation matrix U of the discriminative latent subspace F in which the K clusters are best separated. Following the original idea of Fisher (1936), the functional subspace F should be such that the variance within the groups should be minimum while the variance between

groups should be maximum.

Let $\mathbf{C}$ be the covariance operator of X with kernel

$$C(t,s) = \mathbb{E}\left[(X(t) - m(t))(X(s) - m(s))\right],$$

and $\mathbf{B}$ the integral between cluster covariance operator with kernel

$$B(t,s) = \mathbb{E}\left[\mathbb{E}[X(t) - m(t)|Z]\mathbb{E}[X(s) - m(s)|Z]\right],$$

where $m(t) = \mathbb{E}[X(t)]$. In the following and without loss of generality, the curves are assumed to be centered, *i.e.* $m(t) = 0$. The operator $\mathbf{B}$ can thus be rewritten as:

$$\begin{aligned} B(t,s) &= \mathbb{E}\left[\mathbb{E}[X(t)|Z]\mathbb{E}[X(s)|Z]\right], \\ &= \mathbb{E}\left[\sum_{k=1}^{K} \mathbf{1}_{\{Z=k\}}\mathbb{E}[X(t)|Z=k] \sum_{\ell=1}^{K} \mathbf{1}_{\{Z=\ell\}}\mathbb{E}[X(s)|Z=\ell]\right] \\ &= \sum_{k=1}^{K} P(Z=k)\mathbb{E}[X(t)|Z=k]\mathbb{E}[X(s)|Z=k]. \end{aligned}$$

The Fisher criterion, in the functional case and the supervised setting (Preda, Saporta and Lévéder, 2007), looks for the discriminative function $u \in L_2[0,T]$ which is solution of:

$$\max_u \frac{Var(\mathbb{E}[\Phi(X)|Z])}{Var(\Phi(X)),} \tag{2.7}$$

where $\Phi(X) = \int_{[0,T]} X(t)u(t)dt$ is the projection of X on the discriminative function u. Let us recall that we consider here the unsupervised setting and Z is an unobserved variable. The solution of (2.7) is the eigenfunction u associated with the largest eigenvalue $\eta \in \mathbb{R}$ of the following generalized eigenproblem:

$$\begin{aligned} \mathbf{B}u &= \eta \mathbf{C}u \\ \int_{[0,T]} B(t,s)u(s)ds &= \eta \int_{[0,T]} C(t,s)u(s)ds, \end{aligned} \tag{2.8}$$

under the constraint $< u, \mathbf{C}u >_{L_2[0,T]} = 1$.

The estimator for $C(t,s)$ from the sample $\{x_1, ..., x_n\}$, expanded on the basis $(\psi_j)_{j=1,...,p}$, is:

$$\begin{aligned} \hat{C}(t,s) &= \frac{1}{n}\sum_{i=1}^{n}(\sum_{j=1}^{p}\gamma_{ij}\psi_j(t))(\sum_{j=1}^{p}\gamma_{ij}\psi_j(s)) \\ &= \frac{1}{n}\Psi'(t)\mathbf{\Gamma}'\mathbf{\Gamma}\Psi(s), \end{aligned}$$

where $\mathbf{\Gamma} = (\gamma_{ij})_{i,j}$ is the $n \times p$-matrix of basis expansion coefficients and $\Psi(s)$ is the p-vector of the basis functions $\psi_j(s)$ ($1 \leq i \leq n$ and $1 \leq j \leq p$).

Since the variable Z is unobserved, $B(t,s)$ has to be estimated conditionally on the posterior probabilities $t_{ik}^{(q-1)} = E[z_{ik}|\gamma_i, \theta^{(q-1)}]$ obtained from the E step at iteration $q-1$:

$$\begin{aligned}\hat{B}^{(q)}(t,s) &= \sum_{k=1}^{K} \frac{n_k^{(q-1)}}{n} \left(\frac{1}{n_k^{(q-1)}} \sum_{i=1}^{n} t_{ik}^{(q-1)} x_i(t)\right)\left(\frac{1}{n_k^{(q-1)}} \sum_{i=1}^{n} t_{ik}^{(q-1)} x_i(s)\right) \\ &= \frac{1}{n} \sum_{k=1}^{K} \frac{1}{n_k^{(q-1)}} \left(\sum_{i=1}^{n} t_{ik}^{(q-1)} \sum_{j=1}^{p} \gamma_{ij}\psi_j(t)\right)\left(\sum_{i=1}^{n} t_{ik}^{(q-1)} \sum_{j=1}^{p} \gamma_{ij}\psi_j(s)\right),\end{aligned}$$

and in a matrix form:

$$\hat{B}^{(q)}(t,s) = \frac{1}{n}\Psi'(t)\mathbf{\Gamma}'\mathbf{T}\mathbf{T}'\mathbf{\Gamma}\Psi(s),$$

with $n_k^{(q-1)} = \sum_{i=1}^{n} t_{ik}^{(q-1)}$ and $\mathbf{T} = \left(\frac{t_{ik}^{(q-1)}}{\sqrt{n_k^{(q-1)}}}\right)_{i,k}$ is a $n \times K$-matrix.

Assuming that the discriminative function u can be decomposed in the same basis as the observed curves:

$$u(t) = \sum_{j=1}^{p} \nu_j \psi_j(t) = \Psi'(t)\nu, \tag{2.9}$$

the generalized eigenproblem (2.8) becomes:

$$\int_{[0,T]} \frac{1}{n}\Psi'(t)\mathbf{\Gamma}'\mathbf{T}\mathbf{T}'\mathbf{\Gamma}\Psi(s)\Psi'(s)\nu ds = \eta \int_{[0,T]} \frac{1}{n}\Psi'(t)\mathbf{\Gamma}'\mathbf{\Gamma}\Psi(s)\Psi'(s)\nu ds,$$

which is equivalent to

$$\frac{1}{n}\Psi'(t)\mathbf{\Gamma}'\mathbf{T}\mathbf{T}'\mathbf{\Gamma}W\nu = \eta\frac{1}{n}\Psi'(t)\mathbf{\Gamma}'\mathbf{\Gamma}W\nu,$$

with $\mathbf{W} = \int_{[0,T]} \Psi(s)\Psi'(s)ds$. Since this equality holds for all $t \in [0,T]$, we have

$$\mathbf{\Gamma}'\mathbf{T}\mathbf{T}'\mathbf{\Gamma}W\nu = \eta\mathbf{\Gamma}'\mathbf{\Gamma}W\nu,$$

or equivalently

$$(\mathbf{\Gamma}'\mathbf{\Gamma}W)^{-1}\mathbf{\Gamma}'\mathbf{T}\mathbf{T}'\mathbf{\Gamma}W\nu = \eta\nu. \tag{2.10}$$

Finally, the basis expansion coefficient $\nu = (\nu_1, \ldots, \nu_p)'$ of the discriminative function u is the eigenvector of the above generalized eigenproblem associated to the largest eigenvalue. Once the first discriminative function, let us say u_1, is determined, the second discriminative function is obtained by solving the generalized eigenproblem (2.10) in the complementary space of u_1. This procedure is recursively applied until the d discriminative functions $\{u_1, ..., u_d\}$ are obtained. The basis expansion coefficients $\nu_j^{(q)} = (\nu_{j1}^{(q)}, \ldots, \nu_{jp}^{(q)})'$, $j = 1, \ldots, d$, of the estimated discriminative functions are gathered in the $p \times d$ matrix $U^{(q)} = \left(\nu_{j\ell}^{(q)}\right)_{j,\ell}$.

The M step. Following the classical scheme of the EM algorithm, this step aims at maximizing, conditionally to the orientation matrix $U^{(q)}$ obtained from the previous step, the conditional expectation of the complete data log-likelihood $Q(\theta; \theta^{(q-1)}) = E\left[\ell(\theta; \mathbf{\Gamma}, z_1, ..., z_n) | \mathbf{\Gamma}, \theta^{(q-1)}\right]$:

$$\begin{aligned} Q(\theta; \theta^{(q-1)}) &= -\frac{1}{2} \sum_{k=1}^{K} n_k^{(q-1)} \left[\log|\Sigma_k| + (p-d)\log(\beta) - 2\log(\pi_k) + p\log(2\pi) \right. \\ &\quad \left. + \frac{1}{n_k^{(q-1)}} \sum_{i=1}^{n} t_{ik}^{(q-1)} (\gamma_i - \mu_k)^t U^{(q)} \Delta_k^{-1} U^{(q)t} (\gamma_i - \mu_k) \right] \\ &= -\frac{1}{2} \sum_{k=1}^{K} n_k^{(q-1)} \left[\log|\Sigma_k| + (p-d)\log(\beta) - 2\log(\pi_k) + p\log(2\pi) \right. \\ &\quad \left. + trace(\Sigma_k^{-1} U^{(q)t} C_k U^{(q)}) + \frac{1}{\beta} \left(trace(C_k) - \sum_{j=1}^{d} \nu_j^{(q)t} C_k \nu_j^{(q)} \right) \right], \end{aligned}$$

where $\theta = (\pi_k, \mu_k, \Sigma_k, \beta)_k$, for $1 \leq k \leq K$, and $C_k = \frac{1}{n_k^{(q-1)}} \sum_{i=1}^{n} t_{ik}^{(q-1)} (\gamma_i - \mu_k^{(q-1)})(\gamma_i - \mu_k^{(q-1)})^t$.

The maximization of $Q(\theta; \theta^{(q-1)})$ according to π_k, μ_k, Σ_k and β yields the following updates for model parameters:

- $\pi_k^{(q)} = n_k^{(q-1)}/n$,
- $\mu_k^{(q)} = \frac{1}{n_k^{(q-1)}} \sum_{i=1}^{n} t_{ik}^{(q-1)} U^{(q)t} \gamma_i$,
- $\Sigma_k^{(q)} = U^{(q)t} C_k^{(q)} U^{(q)}$,
- $\beta^{(q)} = \left(\text{trace}(C^{(q)}) - \sum_{j=1}^{d} u_j^{(q)t} C^{(q)} u_j^{(q)} \right) / (p-d)$.

Update formula for other models of the family can be easily obtained from Bouveyron and Brunet (2012).

The E step. This last step reduces to update, at iteration q, the posterior probabilities $t_{ik}^{(q)} = E[z_{ik}|\gamma_i, \theta^{(q)}]$. Let us also recall that $t_{ik}^{(q)}$ is as well the posterior probability $P(z_{ik} = 1|\gamma_i, \theta^{(q)})$ that the curve x_i belongs to the kth component of the mixture under the current model. Using Bayes' theorem, the posterior probabilities $t_{ik}^{(q)}$, $i = 1, ..., n$, $k = 1, ..., K$, can be expressed as follows:

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} \phi(\gamma_i, \theta_k^{(q)})}{\sum_{l=1}^{K} \pi_l^{(q)} \phi(\gamma_i|\theta_l^{(q)})}, \tag{2.11}$$

where $\theta_k^{(q)} = (\pi_k^{(q)}, \mu_k^{(q)}, \Sigma_k^{(q)}, \beta^{(q)})$ is the set of parameters for the kth component updated in the M step.

2.5. *Model selection.* We discuss now both the choice of the most appropriate model with the family and the problem of selecting the number K of groups and the intrinsic dimension d. In our case, the problem of selecting K and d can be in fact recasted as a model selection problem. Since a model is defined by its number of component K and its parametrization, model selection allows to both select a parametrization and a number of components.

Classical tools for model selection include the AIC (Schwarz, 1978) and BIC (Schwarz, 1978) criteria which penalize the log-likelihood $\ell(\hat{\theta})$ as follows, for model $\mathcal{M}$:

$$\text{AIC}(\mathcal{M}) = \ell(\hat{\theta}) - \xi(\mathcal{M}), \;\; \text{BIC}(\mathcal{M}) = \ell(\hat{\theta}) - \frac{\xi(\mathcal{M})}{2} \log(n), \tag{2.12}$$

where $\xi(\mathcal{M})$ is the number of free parameters of the model and n is the number of observations. The value of $\xi(\mathcal{M})$ is of course specific to the model selected by the practitioner (*cf.* Table 1). Although penalized likelihood criteria are widely used and asymptotically consistent, AIC and BIC are also known to be less efficient in practical situations than on simulated cases.

To overcome this drawback in real situations, Birgé and Massart (2007) have recently proposed a data-driven technique, called the "slope heuristic", to calibrate the penalty involving in penalized criteria. The slope heuristic was first proposed in the context of Gaussian homoscedastic least square regression and was then used in different situations, including model-based clustering. Birgé and Massart (2007) proved that it exists a minimal penalty and that considering a penalty equals to twice this minimal penalty allows to approximate the oracle model in term of risk. The minimal penalty is in practice estimated by the slope of the linear part of the log-likelihood $\ell(\hat{\theta})$

with regard to the model complexity. The criterion associated with the slope heuristic is therefore defined by:

$$\mathrm{SHC}(\mathcal{M}) = \ell(\hat{\theta}) - 2\hat{s}\xi(\mathcal{M}), \tag{2.13}$$

where $\hat{s}$ is the slope of the linear part of $\ell(\hat{\theta})$. A detailed overview and advices for implementation are given in Baudry, Maugis and Michel (2012). Section 3 compare the classical model selection criteria with the slope heuristic and the latter will be used in Section 4 on the BSS data.

2.6. *Selection of discriminative basis functions.* A further advantage of the proposed modeling is the possibility of using the discriminative subspace to select the relevant basis functions for discriminating the groups. Indeed, the functional subspace F allows to determine the discriminative basis functions through the loading matrix U which contains the coefficients of the linear relation linking the basis functions with the subspace F. It is therefore expected that basis functions associated with large absolute values of U are particularly relevant for discriminating the groups. An intuitive way to find out the discriminative basis functions would be to keep only large absolute loading variables, by thresholding for instance. Even though this approach is commonly used in practice, it has been particularly criticized by Cadima and Jolliffe (1995) since it induces some misleading information. Here, we propose to select the discriminative basis functions by constraining the optimization problem (2.7) of the F step such that the loading matrix U is sparse (*i.e.*, such that U contains as much zeros as possible). To this end, we follow the approach proposed by Bouveyron and Brunet (2014) which rewrites the constrained Fisher criterion as a ℓ_1-penalized regression problem. We therefore make use of their algorithm (Algorithm 2 of Bouveyron and Brunet, 2014) to maximize the optimization problem (2.7) under ℓ_1-penalization.

3. Numerical experimentations. This section presents some numerical experiments to validate on simulated and benchmark data the approach presented above.

3.1. *Model selection.* We first focus on the problem of model selection. Here, BIC and the slope heuristic are challenged on a set of simulated curves. A sample of $n = 100$ curves is simulated according to the following model,

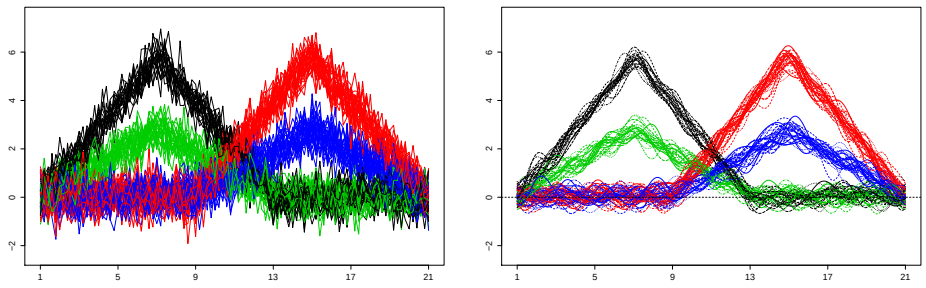

FIG 3.1. *Raw and smoothed simulated curves.*

inspired by Ferraty and Vieu (2003); Preda (2007):

$$\begin{aligned}
\text{Cluster 1}: \quad & X(t) = U + (1-U)h_1(t) + \epsilon(t), \quad && t \in [1,21],\\
\text{Cluster 2}: \quad & X(t) = U + (1-U)h_2(t) + \epsilon(t), \quad && t \in [1,21],\\
\text{Cluster 3}: \quad & X(t) = U + (0.5-U)h_1(t) + \epsilon(t), \quad && t \in [1,21],\\
\text{Cluster 4}: \quad & X(t) = U + (0.5-U)h_1(t) + \epsilon(t), \quad && t \in [1,21],
\end{aligned}$$

where U is uniformly distributed on $[0,1]$ and $\epsilon(t)$ is a white noise, independent from U and such that $\mathbb{V}ar(\epsilon_t) = 0.5$. The function h_1 and h_2 are defined, for $t \in [1,21]$, by $h_1(t) = 6 - |t-7|$ and $h_2(t) = 6 - |t-15|$. The mixing proportions are equal and the curves are observed in 101 equidistant points ($t = 1, 1.2, \ldots, 21$). The functional form of the data is reconstructed using a Fourier basis smoothing with 25 basis functions. Figure 3.1 plots the simulated curves and the smoothed ones.

For each simulated data set, the number K of clusters is estimated thanks to both the BIC criterion and the slope heuristic. As an example of result, Figures 3.2 and 3.3 (right panel) plot respectively the values of the BIC criterion and the slope heuristic for one simulation with the model $\text{DFM}_{[\Sigma_k \beta_k]}$. On this run, both criteria succeed in selecting the actual number of clusters ($K = 4$). Figure 3.3 may require some further explanations. The left panel plots the log-likelihood function with regard to the number of free model parameters, the latter being a function of K (see Table 1). The slope heuristic consists in using the slope of the linear part of the objective function to calibrate the penalty. The linear part is here represented by the red dashed line and was automatically determined using a robust linear regression. The slope coefficient is then used to compute the penalized log-likelihood function, shown on the right panel. We can see here that the slope heuristic provides a penalty close to the one of BIC since both curves are very simi-

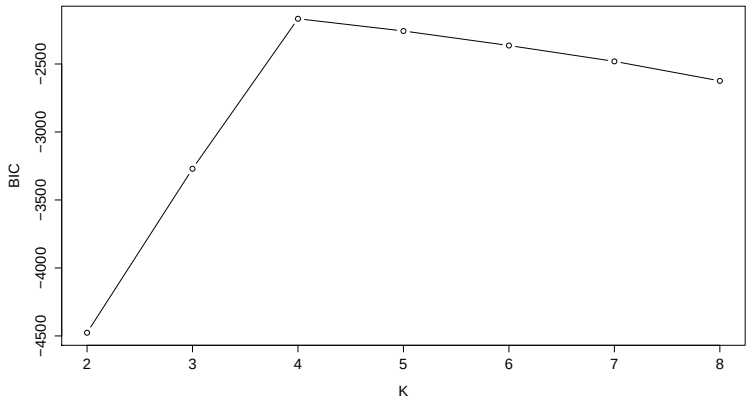

FIG 3.2. *Selection of the number of clusters using BIC on the simulated data (actual value of K is 4).*

lar.

Both criteria were then used to select the appropriate model and number of groups on 100 simulated data sets. Table 2 presents the selected number of clusters by BIC and the slope heuristic over 100 simulations, for each of the 12 DFM models. It turns out that, although BIC can be very efficient when the model is well appropriate, it can provides unsatisfactory results in more difficult inference situations. Conversely, the slope heuristic appear to be more consistent in the selection of the number of clusters while keeping very good overall results. For this reason, the selection of models and number of groups will be addressed in the following section with the slope heuristic.

3.2. *Selection of discriminative basis functions.* This experiment is concerned with the selection of the discriminative basis functions, *i.e.* the most relevant ones for discriminating the clusters. In this work, the selection of the discriminative basis functions is viewed as solving the optimization problem (2.7) of the F step under sparsity constraints (*i.e.*, such that the loading matrix U contains as much zeros as possible). In order to evaluate the ability of our approach to select the relevant discriminative basis functions, we consider now a simulation setting where mainly two different frequencies are involved. The simulation setup is as follows:

$$\begin{aligned}
\text{Cluster 1}: \quad & X(t) = U + (1-U)h_1(t) + \epsilon(t), \qquad t \in [1, 21], \\
\text{Cluster 2}: \quad & X(t) = U + (1-U)h_2(t) + \epsilon(t), \qquad t \in [1, 21], \\
\text{Cluster 3}: \quad & X(t) = U + (1-U)\cos(2t) + \epsilon(t), \qquad t \in [1, 21],
\end{aligned}$$

	Number K of clusters								
Model	2	3	4	5	6	7	8	9	10
$DFM_{[\Sigma_k\beta_k]}$	0	0	**99**	0	0	0	0	1	0
$DFM_{[\Sigma_k\beta]}$	0	0	27	**37**	23	12	1	0	0
$DFM_{[\Sigma\beta_k]}$	0	0	**100**	0	0	0	0	0	0
$DFM_{[\Sigma\beta]}$	0	0	2	2	8	10	10	10	**58**
$DFM_{[\alpha_{kj}\beta_k]}$	0	0	**100**	0	0	0	0	0	0
$DFM_{[\alpha_{kj}\beta]}$	0	0	1	5	8	12	10	7	**57**
$DFM_{[\alpha_k\beta_k]}$	0	0	**100**	0	0	0	0	0	0
$DFM_{[\alpha_k\beta]}$	0	0	0	0	1	1	4	7	**87**
$DFM_{[\alpha_j\beta_k]}$	0	0	**100**	0	0	0	0	0	0
$DFM_{[\alpha_j\beta]}$	0	0	**91**	5	1	1	1	0	1
$DFM_{[\alpha\beta_k]}$	0	0	**100**	0	0	0	0	0	0
$DFM_{[\alpha\beta]}$	0	0	**97**	2	1	0	0	0	0

TABLE 2

Number of clusters selected by BIC over 100 simulations for the 12 DFM models. Actual value for K is 4.

	Number K of clusters								
Model	2	3	4	5	6	7	8	9	10
$DFM_{[\Sigma_k\beta_k]}$	6	9	**84**	0	0	0	0	1	0
$DFM_{[\Sigma_k\beta]}$	15	1	**81**	3	0	0	0	0	0
$DFM_{[\Sigma\beta_k]}$	0	0	**91**	8	1	0	0	0	0
$DFM_{[\Sigma\beta]}$	0	0	**77**	17	5	1	0	0	0
$DFM_{[\alpha_{kj}\beta_k]}$	0	0	**97**	3	0	0	0	0	0
$DFM_{[\alpha_{kj}\beta]}$	0	0	**65**	17	14	3	1	0	0
$DFM_{[\alpha_k\beta_k]}$	0	0	**85**	14	1	0	0	0	0
$DFM_{[\alpha_k\beta]}$	0	0	**78**	14	7	1	0	0	0
$DFM_{[\alpha_j\beta_k]}$	0	1	**87**	11	1	0	0	0	0
$DFM_{[\alpha_j\beta]}$	0	0	**67**	8	6	6	4	3	6
$DFM_{[\alpha\beta_k]}$	4	0	**96**	0	0	0	0	0	0
$DFM_{[\alpha\beta]}$	0	0	**87**	6	4	2	1	0	0

TABLE 3

Number of clusters selected by the slope heuristic over 100 simulations for the 12 DFM models. Actual value for K is 4.

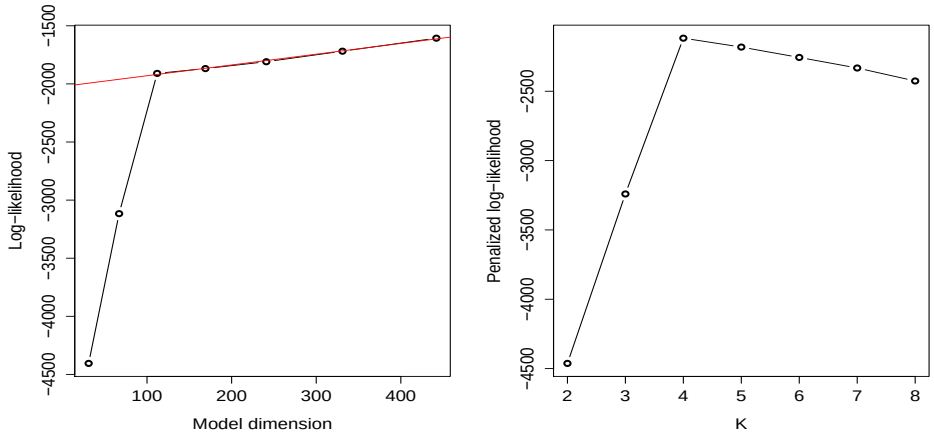

FIG 3.3. *Selection of the number of clusters using the slope heuristic on the simulated data (actual value of K is 4).*

$$\text{Cluster } 4: \qquad X(t) \;=\; U + (1-U)\sin(2t-2) + \epsilon(t), \qquad t \in [1, 21],$$

where U, $\epsilon(t)$, h_1, h_2, the mixing proportions and the observation points are the same as in the previous simulation setting. The functional form of the data is reconstructed using both Fourier basis smoothing (with 25 basis functions) and cubic spline basis (with 50 basis functions). Figure 3.4 plots the simulated curves respectively smoothed on cubic splines and Fourier basis functions.

Starting from the partition estimated with FunFEM and the $\text{DFM}_{[\Sigma_k \beta_k]}$ model, the sparse version of the algorithm is launched with the sparsity parameter $\lambda = 0.1$ on both Fourier and spline smoothed curves. Figures 3.5 and 3.6 plot the selected basis functions on both spline and Fourier bases. For the Fourier basis, the selection of the basis functions indicates which periodicity in the observed curves are the most discriminative, whereas for the spline smoothing, it indicates which time intervals are the most discriminant.

On the one hand, for the Fourier basis, the sparse FunFEM algorithm selects only two discriminative periodicities over the 25 original basis functions (left panel of Figure 3.5). The selected basis functions turn out to be relevant since they actually correspond to the two periodicities present in the simulated data. The right panel of the figure plots the smoothed curves on the two selected basis functions. One can observe that the basis selection is actually relevant since the main features of the data is kept. On the other

hand, for the spline basis, sparse FunFEM has selected three basis functions among the 25 original ones (left panel of Figure 3.6). The three selected functions indicate the most discriminative time intervals. Those time intervals are reported on the right panel of the figure in addition to the curves. One can for instance notice that the first (from the left) selected functions discriminates the green clusters from the three other groups. Similarly, the second discriminative functions allows to separate the black and green clusters from the blue and red curves. Finally, the last selected functions aims at discriminating the black group from the others.

3.3. *Comparison with state-of-the-art methods.* This last numerical study aims at comparing the FunFEM algorithm with state-of-the-art methods on four real data sets: the *Kneading*, *ECG*, *Face*, and *Wafer* data sets.

The Kneading data set (Lévéder et al., 2004) comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process. There are 115 different flours for which the dough resistance is measured during the kneading process for 480 seconds. The data set contains 115 kneading curves observed at 241 equispaced instants of time in the interval $[0, 480]$. The 115 flours produce cookies of different quality: 50 of them have produced cookies of *good* quality, 25 produced *medium* quality and 40 *low* quality. Following (Lévéder et al., 2004; Preda, Saporta and Lévéder, 2007), least square approximation on a basis of cubic B-spline functions (with 18 knots) is used to reconstruct the true functional form of each sample curve. The ECG, Face and Wafer data sets are benchmarks taken from the *UCR Time Series Classification and Clustering* website[1]. The ECG data set consists of 200 electrocardiograms from 2 groups of patients sampled at 96 time instants, and has already been studied in Olszewski (2001). The Face data set (Xi et al., 2006) consists of 112 curves sampled from 4 groups at 350 instants of time. The Wafer data set (Olszewski, 2001) consists of 7174 curves sampled from 2 groups at 152 instants of time. For these three data sets, the same basis of functions as for the Kneading data set has been arbitrarily chosen (20 cubic B-splines).

In this study, FunFEM is compared to the following state-of-the-art methods: k-means with the L_2-metric between curves (d_0) or between their derivatives (d_1) (Ieva et al., 2013) and several model-based methods built by modeling either the functional principal components scores (funclust, Jacques and Preda (2013)) or the basis expansion coefficients: FunHDDC (Bouveyron and Jacques, 2011), fclust (James and Sugar, 2003) and curvclust (Giacofci et al., 2012).

[1] *http://www.cs.ucr.edu/∼eamonn/time_series_data/*

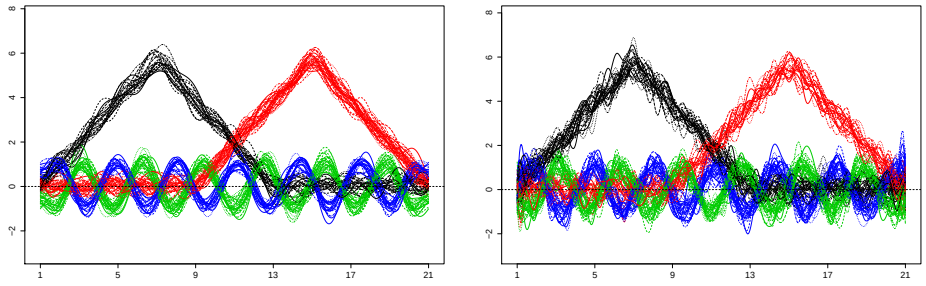

FIG 3.4. *Simulated curves with cubic spline smoothing (left) and Fourier basis smoothing (right).*

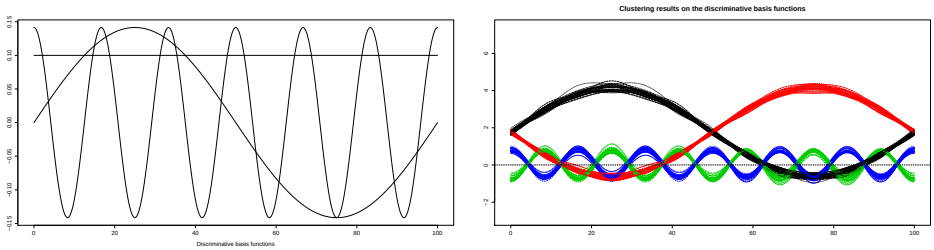

FIG 3.5. *Discriminative functions among the Fourier basis functions: selected basis functions (left) and data projected on the selected basis functions (right).*

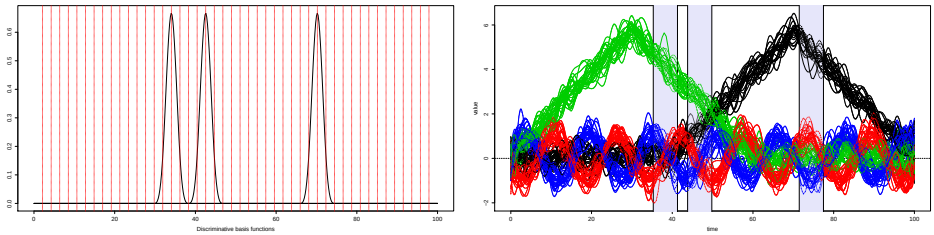

FIG 3.6. *Discriminative functions among the Spline basis functions: selected basis functions (left) and original data with, highlighted in grey, the time periods associated with the selected basis functions (right).*

Method	Kneading	ECG	Face	Wafer
kmeans-d_0	62.61	74.50	48.21	63.34
kmeans-d_1	64.35	61.50	34.80	62.53
Funclust	66.96	**84.00**	33.03	63.10
FunHDDC	62.61	75.00	57.14	63.41
Fclust	64.00	74.50	-	-
Curvclust	65.21	74.50	58.92	63.30
FunFEM DFM$_{[\Sigma_k\beta_k]}$	67.74	71.00	59.82	**66.89**
FunFEM DFM$_{[\Sigma_k\beta]}$	**70.97**	73.00	54.46	64.10
FunFEM DFM$_{[\Sigma\beta_k]}$	67.74	72.00	**61.60**	66.35
FunFEM DFM$_{[\Sigma\beta]}$	66.66	75.00	54.46	64.17
FunFEM DFM$_{[\alpha_{kj}\beta_k]}$	67.74	71.00*	53.57*	**66.89**
FunFEM DFM$_{[\alpha_{kj}\beta]}$	**70.97**	73.50	54.46	64.10
FunFEM DFM$_{[\alpha_k\beta_k]}$	67.74	71.00	53.57	**66.89***
FunFEM DFM$_{[\alpha_k\beta]}$	**70.97**	73.00	57.14	64.10
FunFEM DFM$_{[\alpha_j\beta_k]}$	67.74	72.00	55.35	66.40
FunFEM DFM$_{[\alpha_j\beta]}$	66.66	75.00	53.57	64.17
FunFEM DFM$_{[\alpha\beta_k]}$	67.74*	72.00	53.57	66.40
FunFEM DFM$_{[\alpha\beta]}$	66.66	75.00	56.25	64.17

TABLE 4

Clustering accuracies (in percentage) on the Kneading, Growth, ECG and Wafer data sets for FunFEM and state-of-the-art methods. Bold results correspond to best clustering accuracies and the stars indicate the DFM model selected by BIC.

Table 4 presents the clustering accuracies (according to the known labels) on the four data sets for FunFEM and the six clustering methods. FunFEM turns out to be very competitive with its challengers on those 4 data sets. FunFEM outperforms the other methods on all data sets except on the second one. On the kneading, ECG and wafer sets, the improvement over state-of-the-art methods is significant. It is also worth noticing that the model selected by BIC (the model associated with the higher BIC value) often provides results among the best possible results.

4. Analysis of bike sharing systems. This section now presents the results of the application of FunFEM to data from 8 bike sharing systems (JCDecaux and Transport for London Initiative).

4.1. *The data.* In this work, we used one month of station occupancy data collected on eight bike sharing systems in Europe. The data were collected over 5 weeks, between February, 24 and March, 30, 2014. Table 5 list the BSSs included in this study as well as some information on the systems. The cities were chosen in order to cover different cases in terms of geographic positions of the city (south / north of Europe) and also to cover a range of system sizes, from small scale systems like Nantes to much larger systems

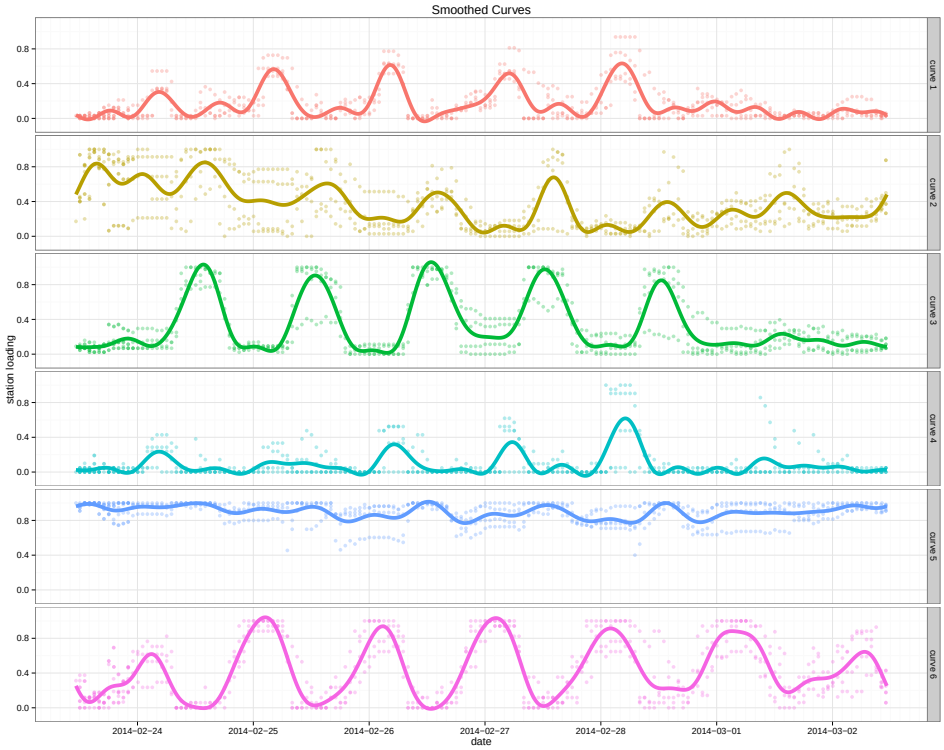

FIG 4.1. *Some examples of smoothed station profiles, with the corresponding observations. One month of observations is depicted here using a period of one week.*

such as Paris.

The station status information, in terms of available bikes and docks, were downloaded every hour during the study period for the seven systems from the open-data API provided by the JCDecaux company [2] and by the Transport for London initiative [3]. To accommodate for the varying stations sizes (in terms of number of docking points) we normalized the number of available bikes by the station size and get a loading profile for each station. The final data set contains 3230 loading profiles, one per station, sampled at 1448 time points, the sampling being not perfectly regular but with an hour in average between two sample points.

The daily and weekly habits of inhabitants introduce a periodic behavior in the BSS station loading profiles, with a natural period of one week. It is

[2]The real time data are available at https://developer.jcdecaux.com/ (with an api key).

[3]The real time data are available at https://www.tfl.gov.uk/info-for/open-data-users/ (with an api key).

City	Stations	Bikes
Paris	1230	18000
London	740	9500
Lyon	345	3200
Bruxelles	330	3800
Valence	280	2400
Seville	260	2150
Marseille	120	650
Nantes	102	880

TABLE 5
Dimension of the eight bike sharing systems involved in the study.

then natural to use a Fourier basis to smooth the curves, with basis functions corresponding to sinus and cosinus functions of periods equal to fractions of this natural period of the data. Using such a procedure the profiles of the 3230 stations were projected on a basis of 41 Fourrier functions. The smoothed curves obtained for 6 different stations are depicted in Figure 4.1, together with the curve samples. The periodic behavior of these profiles is clearly visible in this figure. Some stations exhibit a less clear pattern such as the one of *curve 2* and *curve 5*, but one may clearly observe a week pattern for *curve 1*, *curve 3* and *curve 6*.

4.2. *Clustering results for Paris stations.* We first begin the data analysis with solely the Paris stations. The FunFEM algorithm has been applied on these data with a varying number of clusters, from 2 to 40, and using the $\mathrm{DFM}_{[\alpha_{kj}\beta]}$ model. This model was selected based on the good results it has obtained in the simulation study we performed. To pick an appropriate value for the number of clusters we computed as previously BIC, AIC and the slope heuristic criterion. BIC and AIC provided hard to use values for K since even for 40 clusters they do not reach a maximum. Conversly, the slope heuristic gave a satisfying value for K since it reaches its maximum for $K = 10$. Figure 4.2 shows the evolution of the log-likelihood with respect to the model dimensionality and the associated slope heuristic criterion. On the right panel, the slope heuristic criterion peaks at $K = 10$. This corresponds to an elbow in the log-likelihood function: above this value, the gain in log-likelihood is linear with respect to the model dimensionality. This value of K was used for the cluster analysis. The mean profiles of the obtained clusters are depicted in Figure 4.3, together with the cluster proportions and a sample of curves that belong to each cluster.

The obtained clusters are fairly balanced with around ten percent of the stations each. The clusters are also easily distinguishable. The stations of the first two clusters get bikes during the afternoon and the evening. They

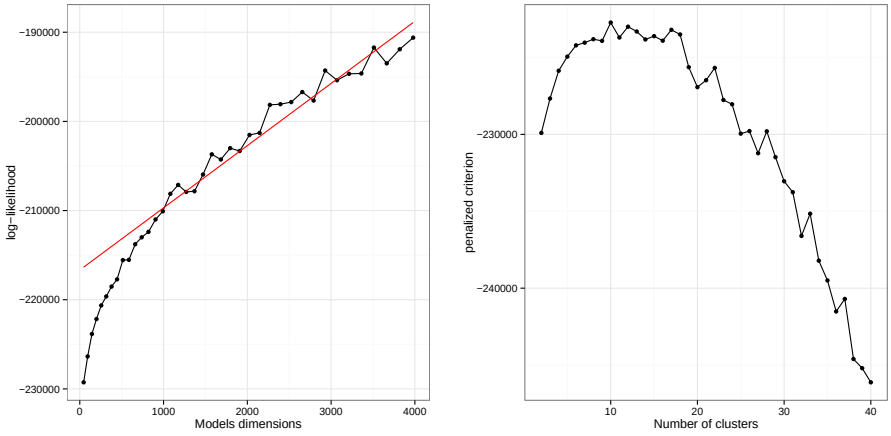

FIG 4.2. *Model selection plots for Paris: log-likelihood with respect to model dimensionality and its estimated linear part (left), slope heuristic criterion with respect to K (right).*

differ during the week-end since the first cluster presents high values during all this period whereas the second cluster experiences a lack of bikes during Saturday mornings. Taking into account these observations, we named the first cluster *Afternoon, Week-end* and the second *Afternoon* as a reference to the periods where these stations are full. The two next clusters present a phase opposition with respect to the previous ones, these stations being full at the end of the morning rush hour (around 9 a.m). As previously these two clusters differ by their week-end behavior, we named the first one *Morning* since these stations are almost empty during all the week-end whereas the second was called *Morning, Week-end* since bikes are available at these stations during a good part of the week-end. The next two clusters do not present the same type of variations, their loading profiles being quite stables during all the week. The difference being in the level of fullness with one cluster around 0.85 of loading and one around 0.7. The first one presents also some day variations which are not visible for the second. We named these clusters *Full* and *Almost Full*.

Clusters 7 and 8 present an overall small activity: cluster 7 get bikes during the night but this does not saturate the stations which reach a balanced state at these time periods. This may correspond to re-allocation journey performed by the operator to balance the system during the night. Cluster 8 oscillates around a balanced state with a little bit more bikes during the afternoons. Taking into account these remarks, we call these clusters *Night rebalancing* and *Balanced*. Lastly, clusters 9 and 10 gather stations which are

almost empty during all the week. Custer 9 presents a quite stable behavior with a constant loading profile around 0.25, whereas the second one smoothly oscillates around 0.1. We call these clusters *Empty* and *Almost empty*.

To complement this analysis of the clustering results, Figure 4.4 presents the spatial location of the clustering results. A first remark which catches the eye in looking at this figure concerns the relatively good spatial organization of the results, even though this information was never used in the clustering process. Stations from the same clusters are quite frequently grouped together on the map. From a Parisian perspective, those results are quite naturals: the *Morning* and *Morning, week-end* clusters (in green on the map) are located in areas with a high employment density which therefore correspond to destinations during the morning commute. This explains why these stations experience a saturation at the end of the morning rush hour. At the opposite, the blue clusters, which correspond to the *Afternooon* and *Afternoon, week-end* clusters, are located in more residential neighborhoods with a higher population density. They therefore correspond to classical origins during the morning rush hour and loose their bikes during this time period. The stations that belong to these clusters are located in regions were they are close from *Empty, Almost empty* stations which are more problematic from a user perspective. These neighborhoods are not in the hyper-center of Paris and they are also located close from stations that belong to the *Night rebalancing* cluster. The *Night rebalancing* cluster is frequently located in up hill locations, such as the "Butte Montmarte" or between the "Père Lachaise" cemetery and the "Butte Chaumont" garden. Lastly, the *Full* and *Almost full* stations are located in the center, whereas the *Balanced* stations are mainly located in the periphery of the system, expect for some stations.

In comparison with previous results obtained with Paris bike share origin/destination data, such as in Côme and Oukhellou (2014), these observations are quite consistent. One of the major difference concerns parks and leisure locations which do not emerge from the clustering in our study. This may be explain by the difference of nature of the input data. The stock data that are used in this paper do not enable the differentiation of these stations whereas origin/destination data allow this. However, stock data are easier to obtain at a large scale and this will allow cross city comparisons, which is the subject of the next section.

4.3. *Clustering results on several cities.* The clustering was also performed on the whole data set which includes stations from the eight systems (see Table 5). The same methodology was used, the curves being projected

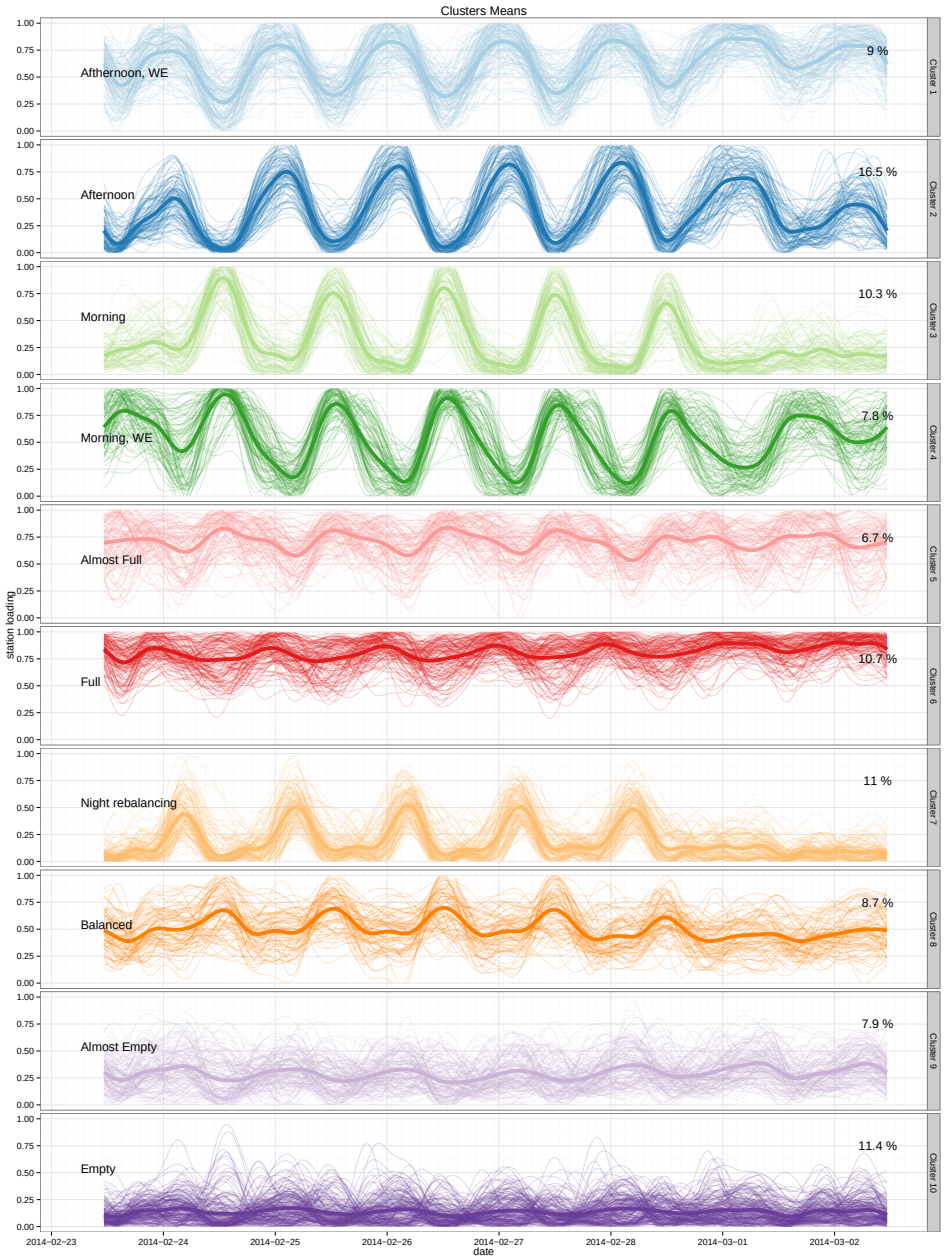

FIG 4.3. *Cluster mean profiles together with 1000 randomly sampled curves. The name of the clusters and their proportions are also provided.*

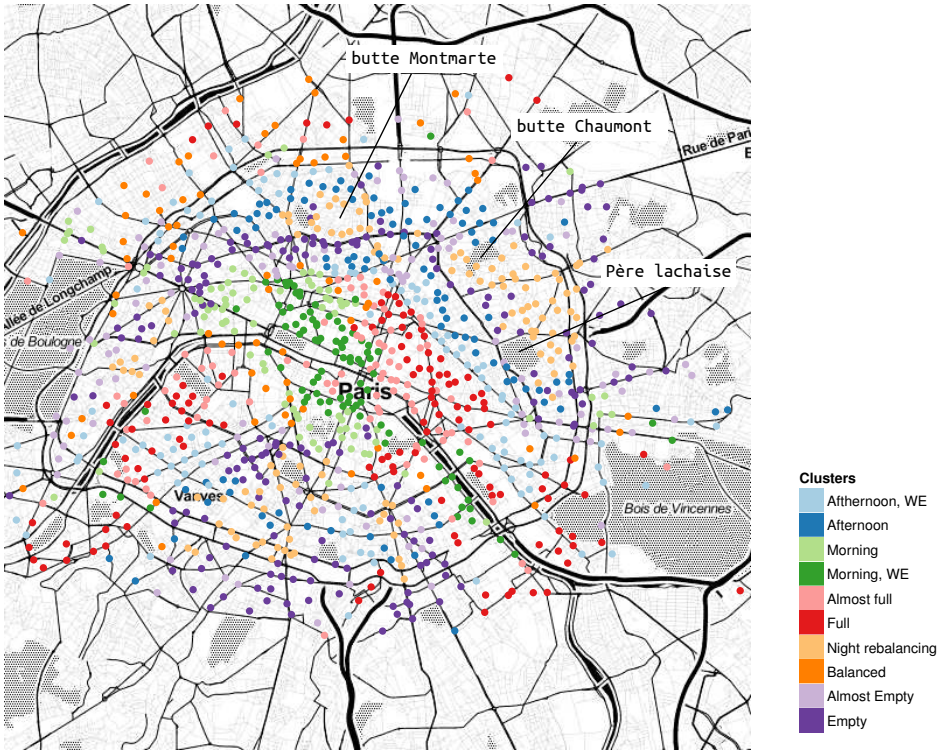

FIG 4.4. *Map of the clustering results. The map background comes from Open-Street data CC-by-SA and Stamen Design.*

on the same Fourier basis and the clustering being performed with the model $\text{DFM}_{[\alpha_{kj}\beta]}$ and with a varying number of cluster from 2 to 40, as previously. The slope heuristic leads to the same number of clusters ($K = 10$ clusters) on this bigger data set. The obtained clusters are also quite close from those obtained on Paris only. Their profiles, which are supplied in Appendix, are close from those shown on Figure 4.3 and their interpretation did not differ significantly. We kept the same labels for the clusters since the main difference come from the amplitude of the profile variations which are smaller on the whole data set. An interesting point in the obtained results concerns the proportions of the different clusters for each cities. This indeed enables an aggregate view of the systems that eases their comparison. These proportions are shown in Figure 4.5.

Some differences between cities are visible on this figure. The proportion of the *Night rebalancing* cluster is for example much important for Paris than in other cities. This cluster, which corresponds to stations which are

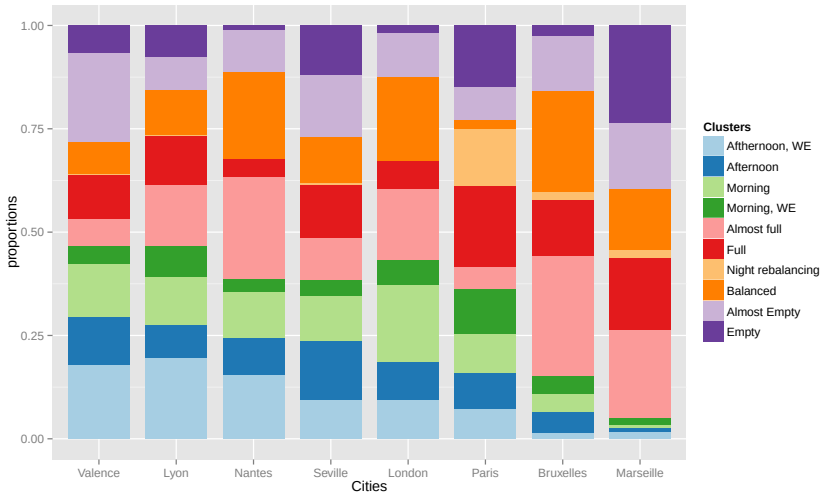

FIG 4.5. *Cluster proportions by city.*

rebalanced during the night, is not visible in the cities other than Paris. At the opposite, the proportions of the *Balanced* cluster is much smaller in Paris than in the other cities. Another clear difference concerns the *Empty* and *Almost empty* clusters stations which are quite important in Marseille and Bruxelles. In Marseille, the *Full* and *Almost full* clusters are also over-represented since they correspond to more than 25% of the city stations. This system seems therefore the more unbalanced system with a lot of stations frequently full or empty. Conversely, the cities on the left of the plot, such as Valence or Lyon, seem to be more active and balanced with an important proportion of stations that belong to the *Afternoon* and *Morning* clusters. This aggregate view allows to identify the BSSs which do not have a satisfying behavior from the exploitation point of view. Indeed, Bruxelles and Marseille have exploitation profiles with a low or even very low proportions of the active clusters (*Afternoon, WE*, *Afternoon*, *Morning, WE* and *Morning*). Conversely, the BSS of Valence, Lyon and London seem the most efficient system. Some of the factors that may explained these behaviors are the ratio between bikes and docks, the topography of the cities together with its geography and the bike redistribution policy.

The observations made on the cluster proportions from Figure 4.5 can be confirmed by looking at the discriminative functional subspace estimated by FunFEM. Figure 4.6 shows the bike stations of the 8 cities projected into the two first axes of the discriminative subspace. The colors indicate

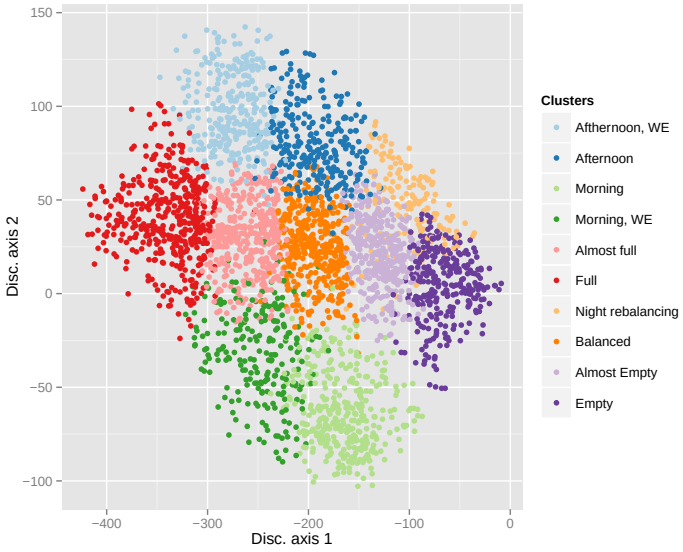

FIG 4.6. *Bike stations projected into the discriminative functional subspace. Colors indicate the cluster belongings.*

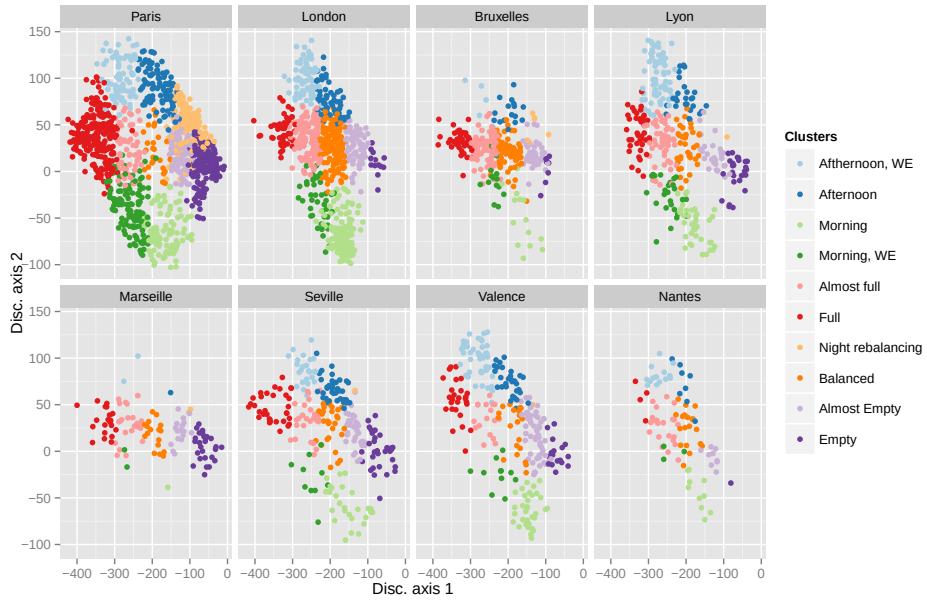

FIG 4.7. *Bike stations per city projected into the discriminative functional subspace. Colors indicate the cluster belongings.*

the cluster belongings of the stations. It may first be useful to interpret the discriminative axes from the cluster meanings. The first axis puts in opposition the *Full* and *Empty* clusters and can be therefore viewed as station loading axis. The second axis opposes the *Afternnon* and *Morning* clusters. It can be therefore linked with the phase of the curves. Figure 4.7 shows the projected bike stations per city on the same discriminative subspace. It appears that most of the studied systems are well represented over the discriminative subspace, except Bruxelles and Marseille. Indeed, those two cities are mainly distributed over the first discriminative axis which opposes the *Full* and *Empty* clusters.

The spatial analysis of the results was also carried out by mapping the clustering results (see Figure A.2 in the appendix). As with Paris, it turns out that the different clusters are also frequently spatially clustered. Furthermore, the same type of global organization is visible for the different cities. Stations from the *Morning* clusters are located in the center of the systems whereas the other cluster are located around the periphery of the system.

5. Conclusion. This work has introduced the discriminative functional mixture (DFM) model which models the data into a discriminative functional subspace. The FunFEM algorithm has also been proposed for the inference of the DFM model. The selection of the most discriminative basis functions can be done afterward by introducing sparsity through a ℓ_1-type penalization. Numerical experiments have demonstrated the efficiency of proposed clustering technique on both simulated and benchmark data. FunFEM appears to be a good challenger to the best state-of-the-art methods. The numerical experiments have also shown the good behavior of the "slope heuristic" for model selection in this context. The application of FunFEM on one month of usage statistics from 8 bike sharing systems has provided insightful analyses and comparisons on the systems. The obtained results were easily interpretable and useful to get a compact representation of the BSS system behavior. Our experiments have demonstrated that such a methodology can be efficiently used to compare several BSSs on a long time period. From an applicative perspective, these results may also be useful to help in designing the redistribution policy of bikes for the studied BSSs.

References.

APUR, (2006). Etude de localisation des stations de vélos en libre service. Rapport. Technical Report No. 349, Atelier Parisien d'Urbanisme.

Baudry, J.-P., Maugis, C. and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing* **22** 455–470.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability theory and related fields* **138** 33–73.

Borgnat, P., Robardet, C., Rouquier, J. B., Parice, A., Fleury, E. and Flandrin, P. (2011). Shared Bicycles in a City: A Signal processing and Data Analysis Perspective. *Advances in Complex Systems* **14** 1–24.

Bouveyron, C. and Brunet, C. (2012). Simultaneous Model-based Clustering and Visualization in the Fisher Discriminative Subspace. *Statistics and Computing* **22** 301–324.

Bouveyron, C. and Brunet, C. (2014). Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. *Computational Statistics* **in press**.

Bouveyron, C., Girard, S. and Schmid, C. (2007). High Dimensional Data Clustering. *Computational Statistics and Data Analysis* **52** 502–519.

Bouveyron, C. and Jacques, J. (2011). Model-based Clustering of Time Series in Group-specific Functional Subspaces. *Advances in Data Analysis and Classification* **5** 281–300.

Büttner, H. J. Mlasowky, Birkholz, T., Groper, D., Fernandez, a. C., G., E. and Banfi, M. (2011). Optmising Bike Sharing in European Cities, A Handbook Technical Report, Intelligent Energy Europe Program (IEE, OBIS projext).

Cadima, J. and Jolliffe, I. (1995). Loadings and correlations in the interpretation of the principal components. *Journal of Applied Statistics* **22** 203–214.

Côme, E. and Oukhellou, L. (2014). Model-based count series clustering for Bike-sharing system usage mining, a case study with the Vélib' system of Paris. *Transportation Research-Part C Emerging Technologies* **22** 88.

De Maio, P. (2009). Bike-sharing: History, Impacts, Models of Provision, and Future. *Journal of Public Transportation* **12** 41–56.

Dell'Olio, L., Ibeas, A. and Moura, J. L. (2011). Implementing bike-sharing systems. In *ICE - Municipal Engineer* **164** 89–101. ICE publishing.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001). *Pattern Classification*, 2. ed. Wiley, New York.

Escabias, M., Aguilera, A. M. and Valderrama, M. J. (2005). Modeling environmental data by functional principal component logistic regression. *Environmetrics* **16** 95–107.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: a nonparametric approach. *Computational Statistics and Data Analysis* **44** 161–173.

Froehlich, J., Neumann, J. and Oliver, N. (2008). Measuring the pulse of the city through shared bicycle programs. In *UrbanSense08* 16–20.

Froehlich, J., Neumann, J. and Oliver, N. (2009). Sensing and Predicting the Pulse of the City through Shared Bicycling. In *21st International Joint Conference on Artificial Intelligence, IJCAI'09* 1420–1426. AAAI Press.

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business and Economic Statistics* **26** 78–89.

Giacofci, M., Lambert-Lacroix, S., Marot, G. and Picard, F. (2012). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics* **in press**.

Heard, N. A., Holmes, C. C. and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical*

Association **101** 18–29. . MR2252430

Ieva, F., Paganoni, A. M., Pigoli, D. and Vitelli, V. (2013). Multivariate functional clustering for the analysis of ECG curves morphology. *Journal of the Royal Statistical Society. Series C. Applied Statistics* **62** 401–418.

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

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

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98** 397–408.

Lathia, N., Saniul, A. and Capra, L. (2012). Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies* **22** 88–102.

Lévéder, C., Abraham, P. A., Cornillon, E., Matzner-Lober, E. and Molinari, N. (2004). Discrimination de courbes de prÈtrissage. In *ChimiomÈtrie 2004* 37–43.

Lin, J. R. and Yang, T. (2011). Strategic design of public bicycle sharing systems with service level constraints. *Transportation Research Part E: Logistics and Transportation Review* **47** 284–294.

Olszewski, R. T. (2001). Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

Preda, C. (2007). Regression models for functional data by reproducing kernel Hilbert spaces methods. *Journal of Statistical Planning and Inference* **137** 829-840.

Preda, C., Saporta, G. and Lévéder, C. (2007). PLS classification of functional data. *Comput. Statist.* **22** 223–235.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*, second ed. *Springer Series in Statistics*. Springer, New York.

Ray, S. and Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **68** 305–332.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

Vogel, P., Greiser, T. and Mattfeld, D. C. (2011). Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences* **20** 514 – 523.

Vogel, P. and Mattfeld, D. C. (2011). Strategic and Operational Planning of Bike-Sharing Systems by Data Mining - A Case Study. In *ICCL* 127–141. Springer Berlin Heidelberg.

Xi, X., Keogh, E., Shelton, C., Wei, L. and Ratanamahatana, C. A. (2006). Fast Time Series Classification Using Numerosity Reduction. In *23rd International Conference on Machine Learning (ICML 2006)* 1033-1040.

APPENDIX A

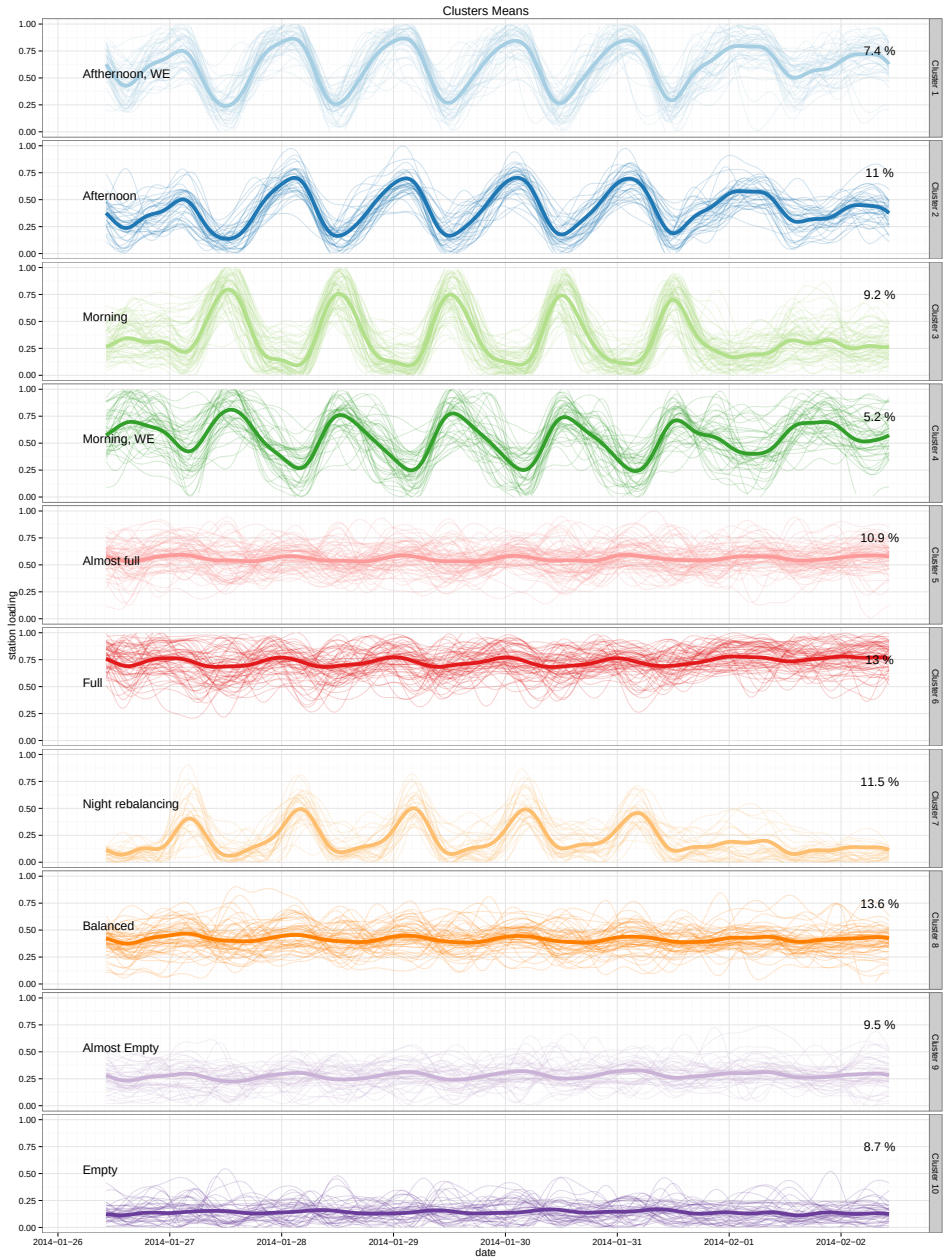

FIG A.1. *Cluster mean profiles together with 1000 randomly sampled curves for the whole data set (Paris, London, Bruxelles, Lyon, Seville and Nantes).*

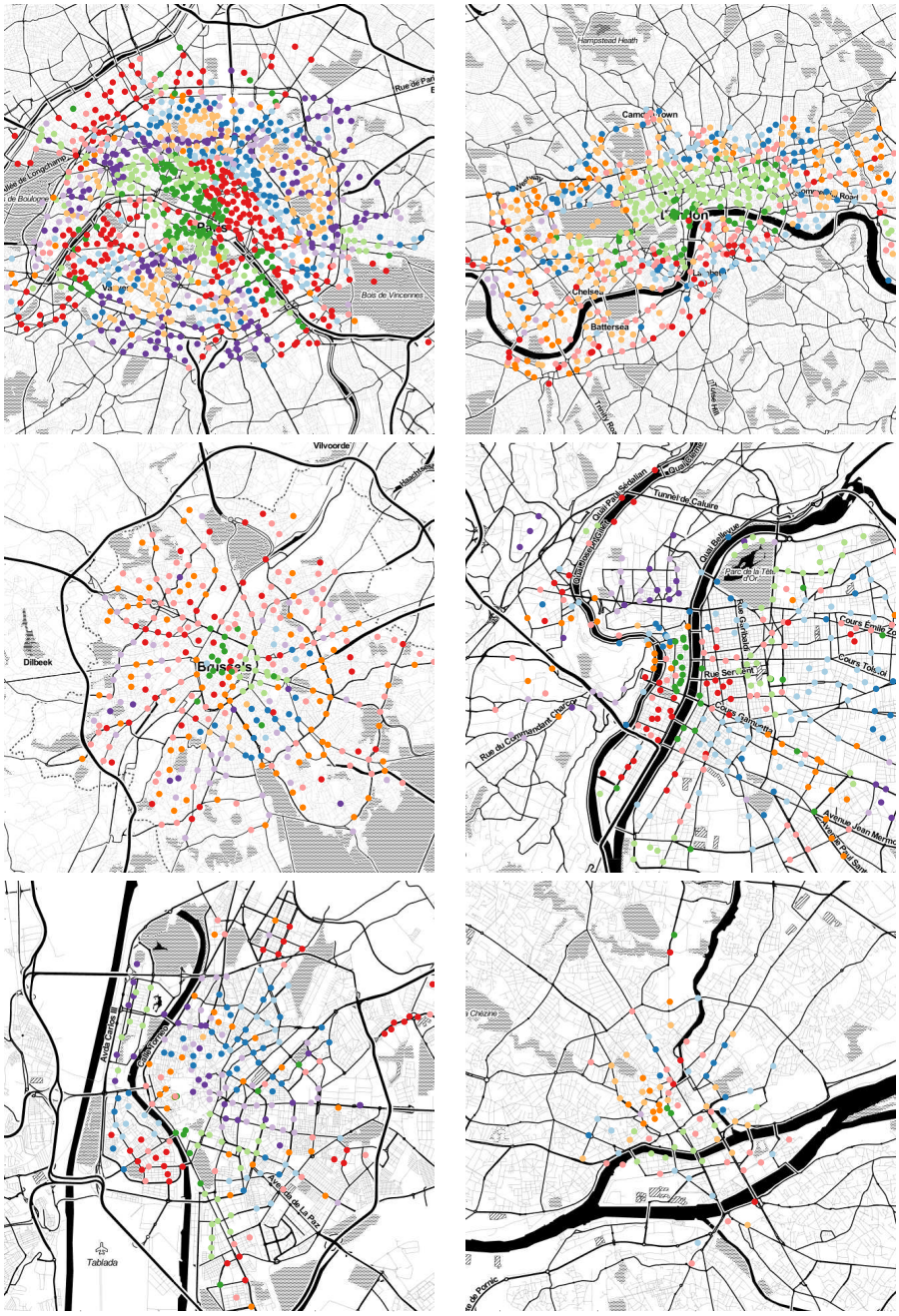

FIG A.2. *Maps of the clustering results (from left to right and top to bottom) for Paris, London, Bruxelles, Lyon, Seville and Nantes. The map background comes from Open-Street data CC-by-SA and Stamen Design.*