Journal of Classification 33:374-411 (2016)
DOI: 10.1007/s00357-016-9212-8

Piecewise Regression Mixture for Simultaneous Functional Data Clustering and Optimal Segmentation

Faicel Chamroukhi

Université de Toulon; Aix Marseille Université; Laboratoire Paul Painlevé, France

Abstract: This paper introduces a novel mixture model-based approach to the simultaneous clustering and optimal segmentation of functional data, which are curves presenting regime changes. The proposed model consists of a finite mixture of piecewise polynomial regression models. Each piecewise polynomial regression model is associated with a cluster, and within each cluster, each piecewise polynomial component is associated with a regime (i.e., a segment). We derive two approaches to learning the model parameters: the first is an estimation approach which maximizes the observed-data likelihood via a dedicated expectation-maximization (EM) algorithm, then yielding a fuzzy partition of the curves into K clusters obtained at convergence by maximizing the posterior cluster probabilities. The second is a classification approach and optimizes a specific classification likelihood criterion through a dedicated classification expectation-maximization (CEM) algorithm. The optimal curve segmentation is performed by using dynamic programming. In the classification approach, both the curve clustering and the optimal segmentation are performed simultaneously as the CEM learning proceeds. We show that the classification approach is a probabilistic version generalizing the deterministic K-means-like algorithm proposed in Hébrail, Hugueney, Lechevallier, and Rossi (2010). The proposed approach is evaluated using simulated curves and real-world curves. Comparisons with alternatives including regression mixture models and the K-means-like algorithm for piecewise regression demonstrate the effectiveness of the proposed approach.

Keywords: Model-based clustering; Functional data analysis; Optimal curve segmentation; Mixture models; Piecewise regression; EM algortihm; CEM algorithm.

We would like to thank the partners of the FUI-SYCIE Project for their financial support to this work.

Corresponding Author's Address: F. Chamroukhi, Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France; Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France; Laboratoire Paul Painlevé, CNRS, UMR 8524, Université Lille 1, 59650 Villeneuve d'Ascq, France, email: faicel.chamroukhi@univ-tln.fr, faicel.chamroukhi@math.univ-lille1.fr.

Published online: 8 November 2016

1. Introduction

Probabilistic modeling approaches are known for their well-established theoretical background and the associated efficient estimation tools in many problems such as regression, classification or clustering. In several situations, such models have interpretation to generalize deterministic algorithms. In this paper, we focus on model-based clustering approaches, that is, the use of mixtures (McLachlan and Peel 2000; Titterington, Smith, and Makov 1985) in cluster analysis. One can cite the following papers, among many others from the broad literature on model-based clustering: Wolfe (1970), Ganesalingam and McLachlan (1978, 1979), McLachlan and Basford (1988), Celeux and Govaert (1995), Fraley and Raftery (2002), Melnykov and Maitra (2010), Samé, Chamroukhi, Govaert, and Aknin (2011), Ingrassia, Minotti, and Vittadini (2012), Lee and McLachlan (2013), Bouveyron and Brunet (2014), Andrews and McNicholas (2014), Lee and McLachlan (2014), Murray, Browne, and McNicholas (2014), Lee and McLachlan (2015), Ingrassia, Punzo, Vittandini, and Minotti (2015), Tang, Browne, and McNicholas (2015), Govaert, Ingrassia, and McLachlan (2015), and Melnykov (2016). In particular, in model-based clustering, that is, the use of mixtures in cluster analysis, for example, the K-means clustering algorithm is well-known to be equivalent to clustering with the Gaussian mixture model (GMM) with the same mixing proportions and identical isotropic covariance matrices when the data are assigned in a hard way after the E-step of the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977; McLachlan and Krishnan 2008), that is, when using the classification expectation-maximization (CEM) algorithm version (see for example Celeux and Govaert 1992, 1993), rather than in a soft way, as in the EM algorithm. We note that the approach based on the CEM algorithm is the same as the so-called classification maximum likelihood as described earlier in McLachlan (1982), and dates back to Scott and Symons (1971).

Most of these statistical analyses in model-based clustering are multivariate, as they involve vectors with reduced dimensionality as the observations (inputs). However, in many application domains, these observations are functions (e.g., curves), and the statistical methods for analyzing such data are functional as they belong to the functional data analysis (FDA) approaches (Ramsay and Silverman 2005). FDA is therefore the paradigm of that data analysis for which the basic unit of information is a function rather than a finite dimensional vector. The flexibility, ease of interpretation, and efficiency of mixture model-based approaches to classification, clustering, segmentation, etc., in multivariate analysis, has lead to a growing investigation for adapting them to the framework of FDA, in particular for curve analysis as in Gaffney and Smyth (1999); Liu and Yang (2009), Gui and

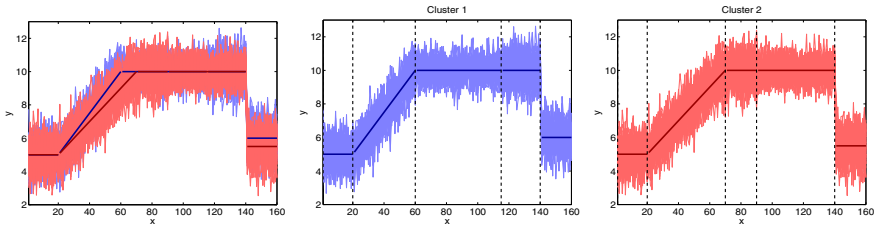

Figure 1. A two-class data set of simulated curves. Each cluster is composed of five noisy constant/linear regimes. The clusters are colored according to the true partition, and the dashed lines represent the true segmentation of each cluster. For the color version of this figure, the reader is referred to the web version of this article.

Li (2003), Shi and Wang (2008), Xiong and Yeung (2004), Chamroukhi, Samé, Govaert, and Aknin (2010), Samé et al. (2011), and Chamroukhi, Hervé, and Samé (2013).

In the present paper, we consider the problem of model-based functional data clustering and segmentation. The considered data are heterogeneous curves which may also present regime changes. The observed curves are univariate and are the values of functions available at discretized input time points. This type of curve can be found in several application domains, including diagnosis application (Chamroukhi, Samé, Govaert, and Aknin (2010), Chamroukhi et al. (2011), bioinformatics (Gui and Li 2003; Picard, Robin, Lebarbier, and Daudin 2007), electrical engineering (Hébrail et al. 2010), etc.

1.1 Problem Statement

Let $((\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n))$ be a set of n independent curves where each curve $(\mathbf{x}_i,\mathbf{y}_i)$ consists of m measurements (observations) $\mathbf{y}_i = (y_{i1},\ldots,y_{im})$ regularly observed at the (time) points $\mathbf{x}_i = (x_{i1},\ldots,x_{im})$ with $x_{i1} < \ldots < x_{im}$. Let $(c_1,\ldots,c_n)$ be the unknown cluster labels of the curves, with $c_i \in \{1,\ldots,K\}$, K being the number of clusters. Figure 1 shows an example from a two-class situation of simulated curves which are mixed at random and each cluster contains five regimes. The aim is to perform curve clustering. As can be seen, each cluster is itself very structured, as it is a succession of non-overlapping segments, which we call regimes. Each regime has its own characteristics and is active for a certain period of time. As can be seen in each of these two clusters, the change of the characteristics of the regimes may include a change in its mean, its variance, its linearity, etc. Thus, in order to precisely infer the hidden structure of the data, it is crucial that the clustering method takes into account the structure of the data which are composed of several regimes, instead of treating them as simple vectors in

$\mathbb{R}^m$. This can be achieved by including a segmentation procedure to capture the various regime changes. This is the regime change problem.

In such a context, basic regression models (e.g., linear, polynomial) are not suitable. The problem of regime changes has been considered as a multiple regime change point detection problem, namely by using Bayesian approaches as in Fearnhead (2006) by using MCMC sampling, and in Fearnhead and Liu (2007) with sequential MCMC for online change point detection. However, these approaches only concern inference from a single curve, and do not concern curve clustering, as they only perform single curve segmentation. An alternative approach in this curve clustering context may consist of using cubic splines to model each class of curves (James and Sugar 2003), but this requires setting the knots a priori. Generative models have been developed by Gaffney and Smyth (1999, 2004), consisting of clustering curves with a mixture of regression models or random effect models. In Liu and Yang (2009), the authors proposed a clustering approach based on random effect regression splines where the curves are represented by B-spline functions. However, the first approach does not address the problem of regime changes and the second one requires setting the spline knots so as learn the model. Another approach based on splines is the clustering sparsely sampled curves in James and Sugar (2003). All these generative approaches use the EM algorithm to estimate the model parameters. Recently, in Hugueney, Hébrail, Lechevallier, and Rossi (2009) and in Hébrail et al. (2010), there has been proposed a distance-based approach based on a piecewise regression model. It allows fitting several constant (or polynomial) models to the curves and performs simultaneous curve clustering and optimal segmentation using a K-means-like algorithm (Hugueney, et al. 2009; Hébrail et al. 2010). The K-means-like algorithm simultaneously performs curve clustering and optimal segmentation using dynamic programming. It minimizes a distance function in the curve space as the learning proceeds. The curve segmentation is carried out using dynamic programming.

The main focus of the present paper is to provide a well-established latent data model to simultaneously perform curve clustering and optimal segmentation. We propose a probabilistic generative model for curve clustering and optimal curve segmentation. It combines both a mixture model to achieve the clustering, and a polynomial piecewise regression model to optimally segment each set (cluster) of homogeneous curves into a finite number of segments using dynamic programming. We show that the proposed probabilistic model generalizes a recently proposed distance-based approach, viz., the K-means-like algorithm of Hébrail et al. (2010). More specifically, the proposed model is a mixture of piecewise regression models. We provide two algorithms for learning the model parameters. The first

one is a dedicated EM algorithm to find a fuzzy partition of the data and an optimal segmentation by maximizing the observed-data log-likelihood. The EM version is the natural way to the maximum likelihood estimation of a mixture model, including the proposed piecewise regression mixture model. The second algorithm maximizes a specific classification likelihood criterion by using a dedicated CEM algorithm in which the curves are partitioned and optimally segmented simultaneously as the learning proceeds. In this CEM-based classification approach, the curves are partitioned in a hard way in contrast to the fuzzy classification approach. The K-means-like algorithm of Hébrail et al. (2010) is shown to be a particular case of the proposed CEM algorithm if some constraints are imposed for the piecewise regression mixture. For the two algorithms, the optimal curve segmentation is performed by using dynamic programming.

This paper is organized as follows. We first briefly recall the two main approaches for model-based clustering, and their extension to curve clustering. Then, Section 2 provides a brief account of related work on model-based curve clustering approaches using polynomial regression mixtures (PRM) and spline regression mixtures (SRM) (Gaffney and Smyth 1999; Gaffney 2004) and recalls the K-means-like algorithm for curve clustering and optimal segmentation based on polynomial piecewise regression (Hébrail et al. 2010). Section 3 introduces the proposed piecewise regression mixture model (PWRM) and its unsupervised learning by deriving both the estimation approach and the classification approach, and the dedicated EM and CEM algorithms. Lastly, Section 6 deals with the experimental study carried out on simulated curves and real-world curves to assess the proposed approach by comparing it to the regression mixtures, the K-means-like algorithm, and the standard GMM clustering.

1.2 Model-Based Clustering

Model-based clustering is the unsupervised classification approach which uses mixture models (McLachlan and Peel 2000; Titterington et al. 1985). Earlier references on the mixture likelihood approach to the unsupervised classification of a sample include, for example, Wolfe (1970), Ganesalingam and McLachlan (1978, 1979), and McLachlan and Basford (1988). One can also cite the pertinent papers of Banfield and Raftery (1993) and Celeux and Govaert (1995) on the use of parsimonious mixtures for clustering, as well in model-based clustering, discriminant analysis, and density estimation in Fraley and Raftery (2002). An account of the subject as well as additional references to the broad literature on the subject can also be found in the survey paper of Melnykov and Maitra (2010) or in the more recent one of Bouveyron and Brunet (2014). Also, the recent special issue edited by Govaert, Ingrassia, and McLachlan (2015) is dedicated to

the subject. Among the very recent contributions to model-based clustering and classification, one can cite the following papers as examples. Andrews and McNicholas (2014) develop an effective variable selection technique for model-based clustering and classification. Bouveyron (2014) contributes a very interesting paper on adaptive mixture discriminant analysis. There has also been a plethora of work on clustering using non-Gaussian mixtures and one can cite some relevant papers, e.g., Lee and McLachlan (2013, 2014); Murray, Browne, and McNicholas (2014), and Lee and McLachlan (2015). Ingrassia et al. (2015) contribute another to a nice series of papers on cluster-weighted models, e.g., Ingrassia, Minotti, and Vittadini (2012). There has also been recent work in other areas including model-based clustering of high-dimensional binary data (Tang, Browne, and McNicholas 2015) and model-based clustering of clickstream data (Melnykov 2016). In the finite mixture approach for cluster analysis, the data probability density function is assumed to be a finite mixture density, each component density being associated with a cluster. The problem of clustering therefore becomes the one of estimating the parameters of the supposed mixture model. In this way, two main approaches are possible, as follows.

1.2.1 The Mixture Approach

In the mixture (or estimation) approach, the parameters of the mixture density are estimated by maximizing the observed-data likelihood. This is generally achieved via the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977; McLachan and Krishnan 2008). After performing the model estimation, the posterior probabilities of the component membership, which represent a fuzzy partition of the data, are then used to determine the cluster memberships through the Bayes allocation rule by assigning each observation to the component (cluster) with the highest posterior probability.

1.2.2 The Classification Approach

The classification approach, also referred to as the maximum likelihood classification approach, consists in optimizing the complete-data likelihood. This maximization can be performed by using the classification version of the EM algorithm, known as the classification EM (CEM) algorithm (Celeux and Govaert 1992). The CEM algorithm inserts a classification step between the E and the M steps of the EM algorithm, which computes the cluster memberships in a hard way by using the Bayes optimal allocation rule.

Note that the outright assignment of the data points to the components of the mixture model after the calculation of the estimate of the parameter vector in the mixture model by acting according to the Bayes' optimal allocation rule was earlier considered by McLachlan (1992, p. 8).

1.3 Model-Based Curve Clustering

Model-based clustering of curves or functional data is a subfield within the broader field of model-based clustering. It is one of several areas of ongoing research in the field of mixture model-based clustering and classification. Mixture model-based curve clustering approaches have been introduced so as to generalize the standard multivariate mixture model to the case of the analysis of functional data where the individuals are presented as curves rather than vectors. Indeed, when the data are curves which are in general very structured, relying on standard multivariate mixture analysis may lead to unsatisfactory results in terms of classification accuracy or modeling accuracy (Chamroukhi, Samé, Govaert, and Aknin 2009b; Chamroukhi 2010; Chamroukhi et al. 2010). However, addressing the problem from the perspective of functional data analysis, that is, formulating functional mixture models, allows overcoming this limitation (Chamroukhi et al. 2009b; Chamroukhi 2010; Chamroukhi et al. 2010; Samé et al. 2011; Gaffney and Smyth 1999; Gaffney 2004; Gaffney and Smyth 2004; Liu and Yang 2009). In the case of model-based curve clustering, one can distinguish the regression mixture approaches (Gaffney and Smyth 1999; Gaffney 2004), including polynomial regression and spline regression, or random effects polynomial regression as in Gaffney and Smyth (2004) or spline regression as in Liu and Yang (2009). Another approach based on splines is concerned with clustering sparsely sampled curves (James and Sugar 2003). All these approaches use the mixture (estimation) approach with the EM algorithm to estimate the model parameters. Another approach, which concerns the mixture-model based clustering of multivariate functional data, is that of Jacques and Preda (2014), in which the clustering is performed in the space of reduced functional principal components. This approach uses an EM-like algorithm.

2. Related Work

In this section, we first describe the model-based curve clustering based on regression mixtures and the EM algorithm (Gaffney 2004; Gaffney and Smyth 2004) as in Chamroukhi et al. (2010) and Chamroukhi et al. (2011). We note that other related papers include Nguyen, McLachlan, and Wood (2015) on regression mixtures for surfaces, as well as Ingrassia, Minotti, and Vittadini (2012) and Ingrassia et al. (2015) on cluster-weighted models. Then we describe the piecewise regression approach to curve clustering and optimal segmentation of Hébrail et al. (2010) and the associated K-means-like algorithm.

2.1 Regression Mixtures and the EM Algorithm for Curve Clustering

Regression mixtures for curve clustering, namely polynomial regression mixture models (PRM) and polynomial spline regression mixtures (PSRM) (Gaffney 2004; Gaffney and Smyth 2004), assume that each curve is drawn from one of K clusters of curves with mixing proportions $(\alpha_1, \ldots, \alpha_K)$. Each cluster of curves is modeled by either a polynomial regression model or a spline regression model. Thus, the mixture density of the ith curve $(i = 1, \ldots, n)$ can be written as

$$p(\mathbf{y}_i|\mathbf{x}_i;\Psi) = \sum_{k=1}^{K} \alpha_k \, \mathcal{N}(\mathbf{y}_i;\mathbf{X}_i\beta_k,\sigma_k^2\mathbf{I}_m), \tag{1}$$

where the α_k's defined by $\alpha_k = p(c_i = k)$ are the mixing proportions, with $\alpha_k > 0$ for each k and $\sum_{k=1}^{K} \alpha_k = 1$; β_k is the coefficient vector and σ_k^2 the noise variance of the kth regression model; and $\mathbf{X}_i$ the design matrix whose construction depends on the adopted model (i.e., polynomial, or polynomial spline, etc). The regression mixture model is therefore fully described by the parameter vector $\Psi = (\alpha_1, \ldots, \alpha_k, \Psi_1, \ldots, \Psi_K)$ with $\Psi_k = (\beta_k, \sigma_k^2)$. The unknown parameter vector Ψ can be estimated by maximizing the observed-data log-likelihood, which is given by

$$\mathcal{L}(\Psi) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \, \mathcal{N}(\mathbf{y}_i;\mathbf{X}_i\beta_k,\sigma_k^2\mathbf{I}_m) \tag{2}$$

via the EM algorithm (Gaffney 2004; Dempster, Laird, and Rubin 1977). The EM algorithm for the regression mixture models and the corresponding updating formula can be found in Gaffney and Smyth (1999) and Gaffney (2004). Once the model parameters have been estimated, a partition of the data into K clusters can then be computed by maximizing the posterior cluster probabilities (Bayes allocation rule).

The regression mixture model however does not address the problem of regime changes within the curves. Indeed, it assumes that each cluster presents a stationary behavior described by a single polynomial mean function. This approach is therefore not well adapted to handle the problem of segmenting curves with regime changes. An alternative is to use polynomial spline regression rather than polynomial regression as in Gaffney (2004), James and Sugar (2003), and Liu and Yang (2009) where the curves are represented by using a combination of several polynomial bases at different time range locations rather than a single polynomial basis. Splines are indeed based on constrained piecewise polynomial fitting with predefined piecewise locations. Therefore, it should be noticed that in spline regression

models, the placement of the knots is generally either fixed by the user or uniform over the range of the input $\mathbf{x}_i$. The optimization of the locations of the knots, which are assumed to be related to the locations of the regime changes (the transition points) in this case of curve segmentation, requires relaxing the regularity constraints for the splines. This leads to the piecewise polynomial regression model (McGee and Carleton 1970; Brailovsky and Kempner 1992; Chamroukhi 2010) in which the placement of the knots can be optimized using dynamic programming (Bellman 1961; Stone 1961).

The piecewise regression model can be used to perform simultaneous curve clustering and optimal segmentation. In Hugueney et al. (2009) and Hébrail et al. (2010), there is proposed a K-means-like algorithm involving a dynamic programming procedure for simultaneous curve clustering and optimal segmentation based on the piecewise regression model. The idea proposed in the present paper is in the same spirit. But it provides a general probabilistic framework to address the problem. In our proposed approach, the piecewise regression model is included in a mixture framework to generalize the deterministic K-means-like approach. Both fuzzy clustering and hard clustering techniques are possible. We note that another possible approach to this task of curve clustering and segmentation is to proceed as in the case of sequential data modeling, in which it is assumed that the observed sequence (in this case a curve) is governed by a hidden process which enables switching from one configuration to another among K configurations. The process usually used in general is a K-state homogeneous Markov chain. This leads to the mixture of hidden Markov models (Smyth 1996) or mixture of hidden Markov model regressions (Chamroukhi et al. 2011).

2.2 Curve Clustering and Optimal Segmentation with K-Means-Like Algorithm

Hébrail et al. (2010) proposes a K-means-like algorithm to simultaneously perform curve clustering and optimal segmentation of each cluster of curves. This is achieved by minimizing a Euclidean distance criterion, as in the standard K-means for multivariate data clustering, while in their functional approach the computations are performed in the space of curves. The curves are partitioned into K clusters and each cluster k is modeled by a piecewise constant regression model and segmented into R_k regimes. The segmentation is performed optimally by using dynamic programming thanks to the additivity of the distance criterion over the set of segments for each cluster. In the following, we recall this technique in order to later show the differences with our proposed approach.

2.2.1 The Optimized Distance Criterion

The clustering and segmentation algorithm proposed in Hugueney et al. (2009) and Hébrail et al. (2010) simultaneously minimizes the following error (distance) criterion:

$$E\left(\mathbf{c},\{I_{kr}\},\{\mu_{kr}\}\right)=\sum_{k=1}^{K}\sum_{i|c_i=k}\sum_{r=1}^{R_k}\sum_{j\in I_{kr}}(y_{ij}-\mu_{kr})^2 \tag{3}$$

with respect to the partition defined by the cluster labels $\mathbf{c}$, and the piecewise cluster parameters $\{\mu_{kr}\}$ and $\{I_{kr}\}$, where $I_{kr}=(\xi_{kr},\xi_{k,r+1}]$ represent the element indices of segment (regime) r ($r=1,\ldots,R_k$) for cluster k and μ_{kr} its constant mean, R_k being the corresponding number of segments. The $m\times 1$ piecewise constant mean curve $\mathbf{g}_k=(g_{k1},\ldots,g_{km})$ where $g_{kj}=\mu_{kr}$ if $j\in I_{kr}$ for all $j=1,\ldots,m$ (i.e., the jth observation y_{ij} belongs to segment r of cluster k) can be seen as the mean curve or the "centroid" of cluster k ($k=1,\ldots,K$). Thus the criterion (3) can be seen as the optimized distortion criterion by the standard K-means for multivariate data clustering, and can then be iteratively minimized by the following K-means-like algorithm (Hébrail et al. 2010).

2.2.2 The K-Means-Like Algorithm

After starting with an initial cluster partition $\mathbf{c}^{(0)}$ (e.g., initialized randomly), the K-means-like algorithm alternates between the two following steps, at each iteration q, until convergence.

Relocation Step. This step consists in finding the optimal piecewise constant prototype for a given cluster k as follows. Based on the current partition $\mathbf{c}^{(q)}$, q being the current iteration number, find the segmentation of each cluster k into R_k regimes by minimizing the following additive criterion:

$$E_k(\mathbf{c}^{(q)},\{I_{kr}\},\{\mu_{kr}\})=\sum_{r=1}^{R_k}\sum_{i|c_i^{(q)}=k}\sum_{j\in I_{kr}}(y_{ij}-\mu_{kr})^2 \tag{4}$$

w.r.t the segment boundaries $\{I_{kr}\}$ and the constant means $\{\mu_{kr}\}$ for each segment. Since (4) is additive over the segments r, the segmentation can be performed in an optimal way by using dynamic programming (Bellman 1961; Stone 1961; Hébrail et al. 2010). Then, each cluster representative is relocated to the piecewise constant prototype $\mathbf{g}_k^{(q)}$ representing the mean of all data points assigned to it.

Assignment Step. This step updates the partition of the curves, $\mathbf{c}$, by assigning each observation $\mathbf{y}_i$ to the nearest piecewise constant prototype $\mathbf{g}_k^{(q)}$ in the sense of the Euclidean distance, that is: $c_i^{(q+1)}=\arg\min_{1\le k\le K}\parallel\mathbf{y}_i-\mathbf{g}_k^{(q)}\parallel^2$.

However, this approach is not probabilistic. It can be seen as deterministic as it does not define a density model on the data. As will be shown in Section 3, it represents a particular case of a more general probabilistic model, the one which we propose. A probabilistic formulation has numerous advantages and relies on a sound statistical background. It is indeed more advantageous to formulate a probabilistic generative approach for ease of interpretation and to help understand the process governing the data generation. In addition, for this clustering task, formulating a latent data model allows considering the clustering naturally within the missing data framework. Furthermore, as we will see, the general probabilistic framework will still be better adapted to the structure of the data, rather than the K-means-like approach which may fail if some constraints on the structure of the data are not satisfied. Another advantage is that the probabilistic approach allows performing soft clustering, which is not generally the case in deterministic approaches. In addition, in probabilistic model-based clustering, we have the possibility of naturally incorporating prior knowledge on the model parameters through prior distributions.

Thus, in the next section we present the proposed piecewise regression mixture model (PWRM) and its unsupervised learning by using two variants of parameter estimation: The first one uses a dedicated EM algorithm and the second one uses a dedicated classification EM (CEM) algorithm. We show how the CEM algorithm used for clustering and optimal segmentation constitutes a probabilistic version of the deterministic approach recalled previously.

3. The Piecewise Regression Mixture (PWRM)

In the proposed approach, the piecewise regression model is stated in a probabilistic framework for model-based curve clustering and optimal segmentation, rather than in a deterministic approach as described previously. First, we present the extension of the standard piecewise regression model for modeling a homogeneous set of independent curves rather than a single curve. Then we derive our piecewise regression mixture model (PWRM).

3.1 Piecewise Regression for Curve Modeling and Optimal Segmentation

As stated in Chamroukhi et al. (2010), piecewise polynomial regression (McGee and Carleton 1970; Brailovsky and Kempner 1992; Ferrari-Trecate and Muselli 2002; Hébrail et al. 2010; Picard et al. 2007) is a modeling and segmentation method that can be used to partition a curve or curves into R regimes (segments). Each segment is characterized by its constant or polynomial mean curve and its variance. The model parameters can be estimated in an optimal way by using a dynamic programming procedure

(Bellman 1961; Stone 1961) thanks to the additivity of the optimized criterion over the regimes (Brailovsky and Kempner 1992; Picard et al. 2007; Hébrail et al. 2010; Hugeuney et al. 2009; Chamroukhi 2010). In the following section, we present the piecewise polynomial regression model, which is generally used for a single curve, in the context of modeling a set of curves. We also describe the algorithm used for parameter estimation by maximizing the likelihood.

3.1.1 Piecewise Regression for Modeling and Optimal Segmentation of a Set of Curves

Piecewise polynomial regression, generally used to model a single curve, (McGee and Carleton 1970; Brailovsky and Kempner 1992; Ferrari-Trecate and Muselli 2002; Chamroukhi, Samé, Govaert, and Aknin 2009a), can be easily used to model a set of curves with regime changes (Chamroukhi et al. 2010; Chamroukhi 2010). The piecewise polynomial regression model assumes that the curves $(\mathbf{y}_1,\ldots,\mathbf{y}_n)$ incorporate R polynomial regimes defined on R intervals $I_1,\ldots,I_R$; the indices of their bounds can be denoted by $\xi=(\xi_1,\ldots,\xi_{R+1})$ where $I_r=(\xi_r,\xi_{r+1}]$ with $\xi_1=0<\xi_2<\ldots<\xi_{R+1}=m$. This defines a partition of the set of curves into R segments of lengths $m_1,\ldots,m_R$: $\{y_{ij}|j\in I_1\},\ldots,\{y_{ij}|j\in I_R\}, i=1,\ldots,n$. The piecewise polynomial regression model for the set of curves, in the Gaussian case, can therefore be defined as follows. For $r=1,\ldots,R$:

$$y_{ij}=\beta_r^\top x_{ij}+\sigma_r\varepsilon_j \quad \text{if } j\in I_r \quad (i=1,\ldots,n; j=1,\ldots,m), \tag{5}$$

where the ε_j are independent zero mean and unit variance Gaussian variables representing additive noise. The model parameters which can be denoted by (θ,ξ) where $\theta=(\beta_1,\ldots,\beta_R,\sigma_1^2,\ldots,\sigma_R^2)$ are composed of the regression parameters and the noise variance for each regime, and are estimated by maximizing the observed-data likelihood. We assume that, given the regimes, the data of each curve are independent. Thus, according to the piecewise regression model (5), the conditional density of a curve is given by:

$$p(\mathbf{y}_i|\mathbf{x}_i;\theta,\xi)=\prod_{r=1}^{R}\prod_{j\in I_r}\mathcal{N}\left(y_{ij};\beta_r^\top x_j,\sigma_r^2\right), \tag{6}$$

and the log-likelihood of the model parameters (θ,ξ) given an independent set of curves $(\mathbf{y}_1,\ldots,\mathbf{y}_n)$ is given by:

$$\begin{aligned}\mathscr{L}(\theta,\xi) &= \log\prod_{i=1}^{n}p(\mathbf{y}_i|\mathbf{x}_i;\theta,\xi)\\ &= -\frac{1}{2}\sum_{r=1}^{R}\left[\frac{1}{\sigma_r^2}\sum_{i=1}^{n}\sum_{j\in I_r}(y_{ij}-\beta_r^T x_{ij})^2+nm_r\log\sigma_r^2\right]+c,\end{aligned} \tag{7}$$

where m_r is the cardinality of I_r (the indices of the points belonging to regime r) and c is a constant term independent of (θ, ξ). Maximizing this log-likelihood is equivalent to minimizing the following criterion

$$\mathscr{J}(\theta,\xi) = \sum_{r=1}^{R}\Big[\frac{1}{\sigma_r^2}\sum_{i=1}^{n}\sum_{j\in I_r}\Big(y_{ij}-\beta_r^T x_{ij}\Big)^2 + nm_r\log\sigma_r^2\Big]. \quad (8)$$

This can be performed by a using dynamic programming procedure thanks to the additivity of the criterion $\mathscr{J}$ over the segments r over the segments (Bellman 1961; Stone 1961). Thus, thanks to dynamic programming, the optimal segmentation can be found. The next section shows how the parameters θ and ξ can be estimated by using dynamic programming to minimize the criterion $\mathscr{J}$ given by (8).

3.1.2 Parameter Estimation of the Piecewise Regression Model by Dynamic Programming

A dynamic programming procedure can be used to minimize the additive criterion (8) with respect to (θ, ξ) or equivalently to minimize the following criterion (9) with respect to ξ:

$$\begin{aligned} C(\xi) &= \min_{\theta}\mathscr{J}(\theta,\xi) \\ &= \sum_{r=1}^{R}\min_{(\beta_r,\sigma_r^2)}\Big[\frac{1}{\sigma_r^2}\sum_{i=1}^{n}\sum_{j=\xi_r+1}^{\xi_{r+1}}\Big(y_{ij}-\beta_k^T x_{ij}\Big)^2 + nm_r\log\sigma_r^2\Big] \\ &= \sum_{r=1}^{R}\Big[\frac{1}{\hat{\sigma}_r^2}\sum_{i=1}^{n}\sum_{j=\xi_r+1}^{\xi_{r+1}}(y_{ij}-\hat{\beta}_r^T x_{ij})^2 + nm_r\log\hat{\sigma}_r^2\Big], \end{aligned} \quad (9)$$

where $\hat{\beta}_r$ and $\hat{\sigma}_r^2$ are the solutions of a polynomial regression problem for segment r and are given by

$$\begin{aligned} \hat{\beta}_r &= \arg\min_{\beta_r}\sum_{i=1}^{n}\sum_{j=\xi_r+1}^{\xi_{r+1}}(y_{ij}-\beta_r^T x_{ij})^2 \\ &= \Big[\sum_{i=1}^{n}\sum_{j=\xi_r+1}^{\xi_{r+1}}x_{ij}x_{ij}^\top\Big]^{-1}\sum_{i=1}^{n}\sum_{j=\xi_r+1}^{\xi_{r+1}}x_{ij}y_{ij}, \end{aligned} \quad (10)$$

and

$$\begin{aligned} \hat{\sigma}_r^2 &= \arg\min_{\sigma_r^2}\frac{1}{\sigma_r^2}\sum_{i=1}^{n}\sum_{j=\xi_r+1}^{\xi_{r+1}}(y_{ij}-\hat{\beta}_r^T x_{ij})^2 + nm_r\log\sigma_r^2 \\ &= \frac{1}{nm_r}\sum_{i=1}^{n}\sum_{j=\xi_r+1}^{\xi_{r+1}}(y_{ij}-\hat{\beta}_r^\top x_{ij})^2. \end{aligned} \quad (11)$$

The matrix form of these solutions can be written as:

$$\hat{\beta}_r = \Big[\sum_{i=1}^{n} \mathbf{X}_{ir}^{\top}\mathbf{X}_{ir}\Big]^{-1} \sum_{i=1}^{n} \mathbf{X}_{ir}\mathbf{y}_{ir}, \tag{12}$$

$$\hat{\sigma}_r^2 = \frac{1}{nm_r}\sum_{i=1}^{n} \|(\mathbf{y}_{ir} - \mathbf{X}_{ir}\hat{\beta}_r)\|^2, \tag{13}$$

where $\mathbf{y}_{ir}$ is the segment (regime) r of the ith curve, that is, the observations $y_{ij}, j = (\xi_r + 1, \ldots, \xi_{r+1})$, and $\mathbf{X}_{ir}$ is its associated design matrix with rows $x_{ij}, j = (\xi_r + 1, \ldots, \xi_{r+1})$ for $i = 1, \ldots, n$.

It can be seen that the criterion $C(\xi)$ given by Equation (9) is additive over the R segments. Thanks to its additivity, this criterion can be optimized globally using a dynamic programming procedure (Bellman 1961; Stone 1961; Brailovsky and Kempner 1992). The piecewise approach provides therefore an optimal segmentation of a homogeneous set of curves into R polynomial segments, each segment being associated with a regime. To handle non-homogeneous sets of curves and at the same time benefit from the efficient segmentation provided by piecewise regression, the model can therefore be integrated in a mixture framework, where each component density will represent a set of curves with a specified number of regimes. This results in the piecewise regression mixture model presented in the next section.

3.2 Piecewise Regression Mixture Model (PWRM) for Curve Clustering and Optimal Segmentation

In this section, we integrate the piecewise polynomial regression model presented previously into a mixture model-based curve clustering framework. Thus, the resulting model is a piecewise regression mixture model which will be abbreviated as PWRM. According to the PWRM model, each curve $(\mathbf{x}_i, \mathbf{y}_i)$ $(i = 1, \ldots, n)$ is assumed to be generated by a piecewise regression model among K models defined by (6), with a prior probability α_k. The distribution of a curve is given by the following piecewise polynomial regression mixture (PWRM) model:

$$p(\mathbf{y}_i|\mathbf{x}_i;\Psi) = \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathcal{N}(y_{ij}; \beta_{kr}^{\top} x_{ij}, \sigma_{kr}^2), \tag{14}$$

where I_{kr} is the set of elements indices of polynomial segment (regime) r for the cluster k, β_{kr} is the $(p+1)$-dimensional vector of its polynomial coefficients, and the α_k are the mixing proportions defined as previously. The parameters of the PWRM model can therefore be denoted by

$$\Psi = (\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K, \xi_1, \ldots, \xi_K),$$

where $\theta_k = (\beta_{k1}, \ldots, \beta_{kR_k}, \sigma^2_{k1}, \ldots, \sigma^2_{kR_k})$ and $\xi_k = (\xi_{k1}, \ldots, \xi_{k,R_k+1})$ are respectively the set of polynomial coefficients and noise variances, and the set of transition points which correspond to the segmentation of the cluster k.

The proposed mixture model is therefore suitable for the clustering and optimal segmentation of complex shaped curves. More specifically, by integrating the piecewise polynomial regression into a mixture framework, the resulting model is able to perform curve clustering. The problem of regime changes within each cluster of curves will be addressed as well thanks to the optimal segmentation provided by dynamic programming for each piecewise regression component model. These two simultaneous outputs are clearly not provided by the standard generative curve clustering approaches, namely the regression mixture and spline regression mixtures. On the other hand, the PWRM is a probabilistic model and as it will be shown in the following, generalizes the deterministic K-means-like algorithm for curve clustering and optimal segmentation.

With the model defined, we now have to estimate its parameters from the data and show how it is used for clustering and optimal segmentation. We present two approaches to estimate the model parameters. The first is an estimation approach and is based on maximizing the observed-data log-likelihood via a dedicated EM algorithm. The second is a classification approach and maximizes the completed-data log-likelihood using a specific CEM algorithm. In the next section we derive the first approach and then we present the second one.

4. Maximum Likelihood Estimation via a Dedicated EM Algorithm

As seen in the Introduction, in the estimation (maximum likelihood) approach, the parameter estimation is performed by maximizing the observed-data (incomplete-data) log-likelihood. Assume we have a set of n i.i.d curves $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ regularly sampled at the time points $\mathbf{x}_i$. According to the model (14), the log-likelihood of Ψ given the observed data can be written as:

$$\mathscr{L}(\Psi) = \log \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{x}_i;\Psi) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \alpha_k \prod_{r=1}^{R_k} \prod_{j \in I_{kr}} \mathscr{N}\left(y_{ij}; \beta_{kr}^{\top} x_{ij}, \sigma^2_{kr}\right). \quad (15)$$

The maximization of this log-likelihood can not be performed in a closed form. The EM algorithm (Dempster, Laird and Rubin 1977; McLachlan and Kirshman 2008) is generally used to iteratively maximize it similarly to the case of standard mixtures. In this framework, the complete-data log-likelihood for a particular partition $\mathbf{c} = (c_1, \ldots, c_n)$, where c_i is the cluster label of the ith curve, is given by

$$\mathscr{L}_c(\Psi,\mathbf{c}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\alpha_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k}\sum_{j\in I_{kr}} z_{ik}\log\mathscr{N}(y_{ij};\beta_{kr}^{\top}x_{ij},\sigma_{kr}^2), \tag{16}$$

where z_{ik} is an indicator binary-valued variable such that $z_{ik}=1$ iff $c_i=k$ (i.e., if the ith curve is generated by cluster k). The next paragraph shows how the observed-data log-likelihood (15) of the proposed model is maximized by the EM algorithm to perform curve clustering and optimal segmentation.

4.1 The EM Algorithm for Piecewise Regression Mixture (EM-PWRM)

The EM algorithm for the polynomial piecewise regression mixture model (EM-PWRM) starts with an initial solution $\Psi^{(0)}$ (e.g., computed from a random partition and uniform segmentation) and alternates between the two following steps until convergence: (e.g., when there is no longer any change in the relative variation of the log-likelihood):

E-Step. The E-step computes the expected complete-data log-likelihood given the observed curves $\mathscr{D}=((\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n))$ and the current value of the model parameters denoted by $\Psi^{(q)}$, q being the current iteration number:

$$\begin{aligned}
&Q(\Psi,\Psi^{(q)})\\
&\quad=\mathbb{E}\left[\mathscr{L}_c(\Psi;\mathscr{D},\mathbf{c})|\mathscr{D};\Psi^{(q)}\right]\\
&\quad=\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left[z_{ik}|\mathscr{D};\Psi^{(q)}\right]\\
&\qquad\qquad\log\alpha_k+\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k}\sum_{j\in I_{kr}}\mathbb{E}\left[z_{ik}|\mathscr{D};\Psi^{(q)}\right]\log\mathscr{N}(y_{ij};\beta_{kr}^{\top}x_{ij},\sigma_{kr}^2)
\end{aligned}$$

$$=\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(q)}\log\alpha_k+\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{R_k}\sum_{j\in I_{kr}}\tau_{ik}^{(q)}\log\mathscr{N}(y_{ij};\beta_{kr}^{\top}x_{ij},\sigma_{kr}^2), \tag{17}$$

where

$$\begin{aligned}
\tau_{ik}^{(q)} &= p(c_i=k|\mathbf{y}_i,\mathbf{x}_i;\Psi^{(q)})\\
&= \frac{\alpha_k^{(q)}\prod_{r=1}^{R_k}\prod_{j\in I_{kr}^{(q)}}\mathscr{N}\left(y_{ij};\beta_{kr}^{T(q)}x_{ij},\sigma_{kr}^{2(q)}\right)}{\sum_{k'=1}^{K}\alpha_{k'}^{(q)}\prod_{r'=1}^{R_{k'}}\prod_{j\in I_{k'r'}^{(q)}}\mathscr{N}(y_{ij};\beta_{k'r'}^{T(q)}x_{ij},\sigma_{k'r'}^{2(q)})}
\end{aligned} \tag{18}$$

is the posterior probability that the ith curve belongs to component k. This step therefore only requires the computation of the posterior cluster probabilities $\tau_{ik}^{(q)}$ $(i = 1, \ldots, n)$ for each of the K clusters.

M-Step. The M-step computes the parameter update $\Psi^{(q+1)}$ by maximizing the Q-function (17) with respect to Ψ, that is:

$$\Psi^{(q+1)} = \arg\max_{\Psi} Q(\Psi, \Psi^{(q)}). \tag{19}$$

To perform this maximization, it can be seen that the Q-function can be decomposed as

$$Q(\Psi, \Psi^{(q)}) = Q_{\alpha}(\alpha_1, \ldots, \alpha_K, \Psi^{(q)}) + \sum_{k=1}^{K} Q_{\Psi_k}\left(\{I_{kr}, \beta_{kr}, \sigma_{kr}^2\}_{r=1}^{R_k}, \Psi^{(q)}\right), \tag{20}$$

where

$$Q_{\alpha}(\alpha_1, \ldots, \alpha_K, \Psi^{(q)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(q)} \log \alpha_r, \tag{21}$$

and

$$\begin{aligned} &Q_{\Psi_k}(\{I_{kr}, \beta_{kr}, \sigma_{kr}^2\}_{r=1}^{R_k}, \Psi^{(q)}) \\ &\quad = \sum_{r=1}^{R_k} \sum_{i=1}^{n} \sum_{j \in I_{kr}} \tau_{ik}^{(q)} \log \mathcal{N}\left(y_{ij}; \beta_{rk}^T x_{ij}, \sigma_{rk}^2\right). \end{aligned} \tag{22}$$

The maximization of $Q(\Psi, \Psi^{(q)})$ can therefore be performed by separate maximizations of Q_{α} (21) with respect to the mixing proportions α_k's and Q_{Ψ_k} (22) with respect to the parameters of each piecewise polynomial regression model $\Psi_k = \{I_{kr}, \beta_{kr}, \sigma_{kr}^2\}_{r=1}^{R_k}$ for $k = 1, \ldots, K$, as follows. The function $Q_{\alpha}(\alpha_1, \ldots, \alpha_K, \Psi^{(q)})$ is maximized with respect to $(\alpha_1, \ldots, \alpha_K) \in [0, 1]^R$ subject to the constraint $\sum_{k=1}^{K} \alpha_k = 1$ using Lagrange multipliers and the updates are given by:

$$\alpha_k^{(q+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(q)}}{n}, \quad (k = 1, \ldots, K). \tag{23}$$

The maximization of (22) corresponds to finding the new update of Ψ_k, that is the piecewise segmentation $\{I_{kr}\}$ of cluster k and the corresponding piecewise regression representation through $\{\beta_{kr}, \sigma_{kr}^2\}$, $(r = 1, \ldots, R_k)$, to the fuzzy cluster k which is composed of the n curves weighted by their posterior probabilities relative to cluster k. Thus, one can observe that each of the maximizations of (22) corresponds to a weighted version of the

piecewise regression problem for a set of curves given by Equation (7), the weights being the posterior cluster probabilities $\tau_{ik}^{(q)}$. Optimizing Q_{Ψ_k} therefore simply consists of solving a weighted piecewise regression problem where the curves are weighted by the posterior cluster probabilities $\tau_{ik}^{(q)}$. The optimal segmentation of each cluster k, represented by the parameters $\{\xi_{kr}\}$ is performed by running a dynamic programming procedure similarly to that of Section 3.1 Equation (9) by weighting the optimization problem. The updating rules for the regression parameters for each cluster of curves correspond to weighted versions of (10) and (11), and are given by

$$\beta_{kr}^{(q+1)} = \Big[\sum_{i=1}^{n} \tau_{ik}^{(q)} \mathbf{X}_{ir}^{\top} \mathbf{X}_{ir}\Big]^{-1} \sum_{i=1}^{n} \mathbf{X}_{ir} \mathbf{y}_{ir}, \tag{24}$$

$$\sigma_{kr}^{2(q+1)} = \frac{1}{\sum_{i=1}^{n} \sum_{j \in I_{kr}^{(q)}} \tau_{ik}^{(q)}} \sum_{i=1}^{n} \tau_{ik}^{(q)} \|(\mathbf{y}_{ir} - \mathbf{X}_{ir} \beta_{kr}^{(q+1)})\|^2, \tag{25}$$

where $\mathbf{y}_{ir}$ is the segment (regime) r of the ith curve, that is, the observations $\{y_{ij} | j \in I_{kr}\}$ and $\mathbf{X}_{ir}$ is its associated design matrix with rows $\{x_{ij} | j \in I_{kr}\}$. Thus, the proposed EM algorithm for the PWRM model provides a fuzzy partition of the curves into K clusters through the posterior cluster probabilities τ_{ik}, each fuzzy cluster being optimally segmented into regimes with indices $\{I_{kr}\}$. When the EM algorithm has converged, a hard partition of the curves can then be deduced by assigning each curve to the cluster which maximizes the posterior probability (18), that is:

$$\hat{c}_i = \arg \max_{1 \leq k \leq K} \tau_{ik}(\hat{\Psi}), \quad (i = 1, \ldots, n). \tag{26}$$

where $\hat{c}_i$ denotes the estimated class label for the ith curve.

To summarize, the proposed EM algorithm computes the maximum likelihood (ML) estimate of the PWRM model. It simultaneously updates a fuzzy partition of the curves into K clusters and an optimal segmentation of each cluster into regimes. At convergence, we obtain the model parameters that include the segments boundaries and the fuzzy clusters. A hard partition of the curves into K clusters is then deduced according to the Bayes' allocation rule by maximizing the posterior probabilities of the component membership.

We note that a similar algorithm for segmentation clustering has been proposed in Picard et al. (2007). This approach uses a dynamic programming procedure with the EM algorithm to segment the temporal gene expression data and the clustering is performed on the segments to assign each set of homogeneous segments to a cluster relative to the spatial behavior of such data. The PWRM model proposed here is quite different from its mixture formulation in the sense that here the curves are supposed to be mixed

at random rather than the segments, so that each cluster is composed of a set of homogeneous temporal curves segmented into heterogeneous segments.

As mentioned in the Introduction, we will propose another scheme to achieve both the model estimation (including the segmentation) and the clustering by using a dedicated Classification EM (CEM) algorithm. In the next section we present the classification approach with its corresponding classification likelihood criterion, and derive the CEM algorithm to maximize it.

5. Maximum Classification Likelihood Estimation via a Dedicated Classification EM Algorithm

The maximum classification likelihood approach simultaneously performs the clustering and the parameter estimation, which includes the curve segmentation, by maximizing the completed-data log-likelihood given by Equation (16) for the proposed PWRM model. The maximization is performed through a dedicated Classification EM (CEM) algorithm.

5.1 The CEM Algorithm for Piecewise Regression Mixture (CEM-PWRM)

The CEM algorithm (Celeux and Govaert 1992) was initially proposed for model-based clustering of multivariate data. We adopt it here in order to perform a model-based curve clustering with the proposed PWRM model. The resulting CEM simultaneously estimates both the PWRM parameters and the classes' labels by maximizing the complete-data log-likelihood given by Equation (16) w.r.t. the model parameters Ψ and the partition represented by the vector of cluster labels $\mathbf{c}$, in an iterative manner as follows. After starting with initial mixture model parameters $\Psi^{(0)}$ (e.g., computed from a randomly chosen partition and a uniform segmentation), the CEM-PWRM algorithm alternates between the two following steps at each iteration q until convergence (e.g., when there is no longer any change in the partition or in the relative variation of the complete-data log-likelihood):

Step 1. The first step updates the cluster labels for the current model defined by $\Psi^{(q)}$ by maximizing the complete-data log-likelihood (16) w.r.t. to the cluster labels $\mathbf{c}$, that is:

$$\mathbf{c}^{(q+1)} = \arg\max_{\mathbf{c}} \mathcal{L}_c(\mathbf{c}, \Psi^{(q)}). \tag{27}$$

Step 2. Given the estimated partition defined by $\mathbf{c}^{(q+1)}$, the second step updates the model parameters by maximizing the complete-data log-likelihood w.r.t. to the PWRM parameters Ψ:

$$\Psi^{(q+1)} = \arg\max_{\Psi} \mathscr{L}_c(\mathbf{c}^{(q+1)}, \Psi). \tag{28}$$

Equivalently, the CEM algorithm therefore consists in inserting a classification step (C-step) between the E- and the M- steps of the EM algorithm presented previously. In the case of the proposed PWRM model, the dedicated CEM-PWRM algorithm runs as follows. It starts with initial model parameters $\Psi^{(0)}$ and then alternates between the three following steps at each iteration q until convergence.

E-Step. The E-step computes the posterior probabilities $\tau_{ik}^{(q)}$ $(i = 1, \ldots, n)$, given by Equation (18), that the ith curve belongs to cluster k, for $i = 1, \ldots, n$ and for each of the K clusters.

C-Step. The C-step computes a hard partition of the n curves into K clusters by estimating the cluster labels through the Bayes allocation rule:

$$c_i^{(q+1)} = \arg\max_{1 \le k \le K} \tau_{ik}^{(q)} \ (i = 1, \ldots, n). \tag{29}$$

M-Step. The M-step, given the estimated cluster labels $\mathbf{c}^{(q+1)}$, updates the model parameters by computing the parameter vector $\Psi^{(q+1)}$ which maximizes the complete-data log-likelihood (16) with respect to Ψ. By rewriting the complete-data log-likelihood given the current estimated partition as

$$\begin{aligned}\mathscr{L}_c(\Psi, \mathbf{c}^{(q+1)}) &\\ = \sum_{k=1}^{K} \sum_{i|c_i^{(q)}=k} \log \alpha_k + \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{i|c_i^{(q)}=k} \sum_{j \in I_{kr}} \log \mathscr{N}(y_{ij}; \beta_{kr}^{\top} x_{ij}, \sigma_{kr}^2),&\end{aligned} \tag{30}$$

we can see that this function can be optimized by separately optimizing the two terms of the r.h.s. of (30). More specifically, the mixing proportions α_k's are updated by maximizing the function $\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(q+1)} \log \alpha_k$ w.r.t. $(\alpha_1, \ldots, \alpha_K) \in [0,1]^K$ subject to the constraint $\sum_{k=1}^{K} \alpha_k = 1$. This is performed by using Lagrange multipliers and gives the following updates:

$$\alpha_k^{(q+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ik}^{(q)} \quad (k = 1, \ldots, K). \tag{31}$$

The regression parameters and the segmentation, which are denoted by $\{\Psi_k\}$ $= \{(\theta_k, \xi_k)\}$ for each of the K clusters, are updated by maximizing the second term of the r.h.s. of (30) similarly as in the case of the EM-PWRM algorithm presented in the previous section. The only difference is that the posterior probabilities τ_{ik} in the case of the EM-PWRM algorithm are replaced by the cluster label indicators z_{ik} when using the CEM-PWRM, the

curves being assigned in a hard way rather than in a soft way. This step consists therefore in estimating a piecewise polynomial regression model for the set of curves of each of the K clusters separately. Each polynomial regression model estimation for each cluster of curves is performed using a dynamic programming procedure as in seen in Section 3.1.

5.2 The CEM-PWRM Algorithm as a Generalization the K-Means-Like Algorithm

In this section we show how the proposed PWRM estimated by the CEM algorithm provides a general framework for the K-means-like algorithm of Hébrail et al. (2010) seen in Section 2.2.

Proposition 5.2.1 *The complete-data log-likelihood (16) optimized by the proposed CEM algorithm for the piecewise regression mixture model is equivalent to the distance criterion (3) optimized by the K-means-like algorithm of Hébrail et al. (2010) if the following constraints are imposed:*

- $\alpha_k = \frac{1}{K}\ \forall K$ *(identical mixing proportions)*
- $\sigma_{kr}^2 = \sigma^2\ \forall r = 1,\ldots,R_k$ *and* $\forall k = 1,\ldots,K$ *(isotropic and homoskedastic model)*
- *piecewise constant approximation of each segment of curves rather than a polynomial fitting.*

Therefore, the proposed CEM algorithm for piecewise polynomial regression mixture is the probabilistic version for hard curve clustering and optimal segmentation of the K-means-like algorithm (c.f., Section 2.2). It has a better ability to handle data with nonequivalent population proportions and easily takes into accounts within-cluster variances.

Proof. The complete data log-likelihood (16) can be rewritten as

$$\mathscr{L}_c(\Psi,\mathbf{c}) = \sum_{k=1}^{K}\sum_{i|c_i=k}\log\alpha_k - \frac{1}{2}\sum_{k=1}^{K}\sum_{i|c_i=k}\sum_{r=1}^{R_k}\sum_{j\in I_{kr}}\Big[\Big(\frac{y_{ij}-\boldsymbol{\beta}_{kr}^{\top}x_{ij}}{\sigma_{kr}^2}\Big)^2 + \log(2\pi\sigma_{kr}^2)\Big]. \quad (32)$$

Now, if we consider the constraints in Proposition 5.2.1 for the proposed PWRM model, the maximized complete-data log-likelihood takes the following form:

$$\mathscr{L}_c(\Psi,\mathbf{c}) = \sum_{k=1}^{K}\sum_{i|c_i=k}\log\frac{1}{K} - \frac{1}{2}\sum_{k=1}^{K}\sum_{i|c_i=k}\sum_{r=1}^{R_k}\sum_{j\in I_{kr}}\Big[\Big(\frac{y_{ij}-\mu_{kr}}{\sigma^2}\Big)^2 + \log(2\pi\sigma^2)\Big]. \quad (33)$$

Maximizing this function is therefore equivalent to minimizing the following criterion w.r.t the cluster labels $\mathbf{c}$ and the segments indices I_{kr} and the segments' constant means μ_{kr}:

$$\mathscr{J}\left(\mathbf{c},\{\mu_{kr},I_{kr}\}\right)=\sum_{k=1}^{K}\sum_{r=1}^{R_k}\sum_{i|c_i=k}\sum_{j\in I_{kr}}\left(y_{ij}-\mu_{kr}\right)^2, \tag{34}$$

which is exactly the distortion criterion optimized by the K-means-like algorithm of Hébrail et al. (2010) (cf. Equation (3)).
■

5.3 Model Selection

The problem of model selection here is equivalent to choosing the optimal number of clusters K, number of regimes R, and polynomial degree p. The optimal value of the triplet (K,R,p) can be computed by using some model selection criteria such as the Bayesian Information Criterion (BIC) (Schwarz 1978), as in Liu and Yang (2009) or the Integrated Classification Likelihood criterion (ICL) (Biernacki, Celeux and Govaert 2000), etc. Let us recall that the BIC is a penalized log-likelihood criterion which can be defined as a function to be maximized that is given by: $\mathrm{BIC}(K,R,p)=\mathscr{L}(\hat{\Psi})-\frac{\nu_\Psi\log(n)}{2}$, whereas the ICL is a penalized complete-data log-likelihood and can be expressed as follows: $\mathrm{ICL}(K,R,p)=\mathscr{L}_c(\hat{\Psi})-\frac{\nu_\Psi\log(n)}{2}$, where $\mathscr{L}(\hat{\Psi})$ and $\mathscr{L}_c(\hat{\Psi})$ are respectively the incomplete (observed) data log-likelihood and the complete data log-likelihood, obtained at convergence of the (C)EM algorithm, $\nu_\Psi=\sum_{k=1}^{K}R_k(p+3)-1$ is the number of free parameters of the model and n is the sample size. The number of free model parameters includes $K-1$ mixing proportions, $\sum_{k=1}^{K}R_k(p+1)$ polynomial coefficients, $\sum_{k=1}^{K}R_k$ noise variances and $\sum_{k=1}^{K}(R_k-1)$ transition points.

6. Experimental Study

In this section, we assess the proposed PWRM with both the EM and CEM algorithms in terms of curve clustering and segmentation. We study the performance of the developed PWRM model by comparing it to the polynomial regression mixture models (PRM) (Gaffney 2004), the standard polynomial spline regression mixture model (PSRM) (Gaffney 2004; Gui and Li 2003; Liu and Yang 2009) and the piecewise regression model used with the K-means-like algorithm (Hébrail et al. 2010). We also include comparisons with standard model-based clustering of multivariate data using Gaussian mixture models (GMM). For all the compared generative approaches we consider both the EM and the CEM algorithms. Thus, the ten compared approaches can be summarized as follows: EM-GMM, EM-PRM,

EM-PRM, EM-PSRM, K-means-like, EM-PWRM and CEM-PWRM. All algorithms have been implemented in Matlab. The aim of including the standard multivariate data clustering with Gaussian mixtures models and the EM algorithm is to show that it is necessary to adapt them to curve clustering approaches as they do not take into account the functional structure of the data. The algorithms are evaluated using experiments conducted on both synthetic and real curves.

6.1 Evaluation Criteria

The algorithms are evaluated in terms of curve classification and approximation accuracy. The evaluation criteria used are the classification error rate between the true simulated partition and the estimated partition, and the intra-cluster inertia $\sum_{k=1}^{K}\sum_{i|\hat{c}_i=k}\|\mathbf{y}_i-\hat{\mathbf{y}}_k\|^2$, where $\hat{c}_i$ indicates the estimated class label of the ith curve from the sample and $\hat{\mathbf{y}}_k=(\hat{y}_{kj})_{j=1,\ldots,m}$ is the estimated mean curve of cluster k. Each point of the mean curve of cluster k is given by

- $\hat{y}_{kj}=\hat{\boldsymbol{\beta}}_{kr}^{\top}x_{ij}$ if $j\in\hat{I}_{kr}$ for the proposed approach (EM-PWRM, CEM-PWRM) and the K-means-like approach of Hébrail et al. (2010),
- $\hat{y}_{kj}=\hat{\boldsymbol{\beta}}_{k}^{\top}x_{ij}$ for both the polynomial regression mixture (PRM) and the spline regression mixtures (PSRM),
- $\hat{y}_{kj}=\frac{\sum_{i=1}^{n}\hat{z}_{ik}y_{ij}}{\sum_{i=1}^{n}\hat{z}_{ik}}$ for the standard model-based clustering with GMM.

6.2 Experiments with Simulated Curves

6.2.1 Simulation Protocol and Algorithms Setting

The simulated data consisted of curves generated from a mixture of two classes, each class being simulated as a piecewise linear function corrupted by Gaussian noise. More specifically, the simulated curves consisted of $n=100$ curves of $m=160$ regularly sampled observations at the discrete time points $\mathbf{t}=(1,\ldots,m)$. The curves are mixed in proportion randomly with mixing proportions α_k, $(k=1,2)$. We first considered uniform mixing proportions ($\alpha=[0.5,0.5]$) and then varied the proportions between the two classes so as to have non-uniformly mixed classes. In the simulated curves, we consider variation in mean, variance, and regime shape (constant, linear). Table 1 shows the simulation parameters used to generate each observation $\mathbf{y}_i=(y_{ij})_{j=1}^{m}$ and Figure 1 shows an example of simulated curves for this situation.

6.2.2 Algorithms Setting

The algorithms are initialized from a random partition for the clustering. For the segmentation, the models performing segmentation are ini-

Table 1. Simulation parameters: σ_{kr} represents the noise standard deviation for regime r of cluster k, ξ_k the transition points within cluster k, and $e_j \sim \mathcal{N}(0,1)$ are zero-mean unit-variance Gaussian variables representing an additive noise.

regime	cluster $k=1$		cluster $k=2$	
r=1	$[5+\sigma_{11}e_{ij}]\mathbb{1}_{[1,20]}$	$\sigma_{11}=0.8$	$[5+\sigma_{11}\ e_{ij}]\mathbb{1}_{[1,20]}$	$\sigma_{21}=0.8$
r=2	$[0.125j+2.5+\sigma_{12}e_{ij}]\mathbb{1}_{]20,60]}$	$\sigma_{12}=0.8$	$[0.1j+3+\sigma_{22}e_{ij}]\mathbb{1}_{]20,70]}$	$\sigma_{22}=0.8$
r=3	$[10+\sigma_{13}e_{ij}]\mathbb{1}_{]60,115]}$	$\sigma_{22}=0.6$	$[10+\sigma_{23}e_{ij}]\mathbb{1}_{]70,90]}$	$\sigma_{22}=0.8$
r=4	$[10+\sigma_{14}e_{ij}]\mathbb{1}_{]115,140]}$	$\sigma_{22}=0.8$	$[10+\sigma_{24}e_{ij}]\mathbb{1}_{]90,140]}$	$\sigma_{22}=0.6$
r=5	$[6+\sigma_{15}e_{ij}]\mathbb{1}_{]140,160]}$	$\sigma_{22}=0.8$	$[5.5+\sigma_{25}e_{ij}]\mathbb{1}_{]140,160]}$	$\sigma_{22}=0.8$
	$\xi_1=[1,20,60,115,140,160]$		$\xi_2=[1,20,70,90,140,160]$	

Table 2. Intra-class inertia for the simulated curves

EM-GMM	EM-PRM	EM-PSRM	K-means-like	EM-PWRM	CEM-PWRM
19639	25317	21539	17428	17428	17428

tialized from random contiguous segmentations, including a uniform segmentation. The algorithms are stopped when the relative variation of the optimized criterion between two iterations is less than a predefined threshold (10^{-6}). For the same model parameters, the results are computed for 20 different data sets, and for each data set, we performed 10 runs of each algorithm EM and the solution providing the best value of the optimized criterion was chosen. Each run of the EM algorithm is initialized using the best partition selected from 20 repeated runs of standard K-means, that is, the partition corresponding to the lowest distortion. The obtained clusters of curves are then segmented randomly to initialize the regression parameters. Note that here the number of EM repetitions (10) was chosen as in Section 4.1 of Biernack, Celeux and Govaert (2003), where the authors performed $x = 10$ repetitions. The number of K-means repetitions (20) for each EM run is standard in EM clustering using mixtures and has been revealed to be sufficient for our experiments. The initial guess of parameters provides quite stable EM solutions. For the use of only K-means in clustering, the reader can see Steinley and Brusco (2007) for further details on strategies to obtain stable solutions. For our context of clustering using mixtures and the EM algorithm, the reader is referred to the paper of Biernacki, Celeux and Govaert (2003) on strategies for choosing starting values for the EM algorithm.

6.2.3 Obtained Results

We applied the different models to the simulated curves, where for the piecewise regression model we trained it with linear polynomial regimes ($p = 1$). The polynomial regression mixture (PRM) was trained with a

polynomial degree $p = 10$. For the polynomial spline regression mixture (PSRM), we used cubic splines (of degree $p = 3$) with 20 uniformly placed internal knots. In terms of the numerical results, Table 2 gives the obtained intra-cluster inertias. For this situation, which is extremely difficult, all the algorithms retrieved the actual partition (misclassification error of 0% for all the algorithms). However, in terms of curve approximation, we can clearly see that, on the one hand, the standard model-based clustering using the GMM is not suited, as it does not take into account the functional structure of the curves and therefore does not take into account the smoothness, they rather compute an over-all mean curve. On the other hand, the proposed probabilistic approach (EM-PWRM, CEM-PWRM) and that of Hébrail et al. (2010) (which we denoted here by K-means-like), as expected, provide the same results in terms of clustering and segmentation. This is to be attributed to the fact that the K-means PWRM approach is a particular case of our probabilistic approach. Figure 2 shows the different clustering and segmentation results for the simulated curves given in Figure 1. It can be seen that the best curve approximation are provided by the PWRM models. The GMM mean curves are simply over-all means, and the PRM and the PSRM models, as they are based on continuous curve prototypes, do not take into account the segmentation, in contrast to the PWRM models which are well adapted to perform simultaneous curve clustering and segmentation. We note that in all the experiments, we included both the EM and the CEM algorithm and the results are not significantly different. Hence, we chose to give the results for only one of these two algorithms.

In the previous situation, the algorithms were mainly evaluated regarding the curve approximation while keeping the clustering task not very difficult. Now, we vary the noise level in order to assess the models in terms of curve clustering. This is performed by computing the misclassification error rate for different noise levels. The curves were still simulated according to the same parameters of Table 1 while varying the noise level for all the regimes by adding a noise level variation s to the standard deviation σ_{kr}.

Figure 3 shows the misclassification error rates obtained for the different noise levels.For small variations in the noise level, the results are very similar and comparable to those presented previously. However, as the variation in the noise level increases, the misclassification error rate increases faster for the other models than for the proposed PWRM model. The EM and the CEM algorithm for the proposed approach provide very similar results with a slight advantage for the CEM version.

For the previous situations, the data was simulated according to a mixture with equal mixing proportions. Now we vary the parameters in order to make the mixture with non-uniform mixing proportions ($\alpha_1 = 0.2$ $\alpha_2 = 0.8$) and with a variance change less pronounced than before (namely we set

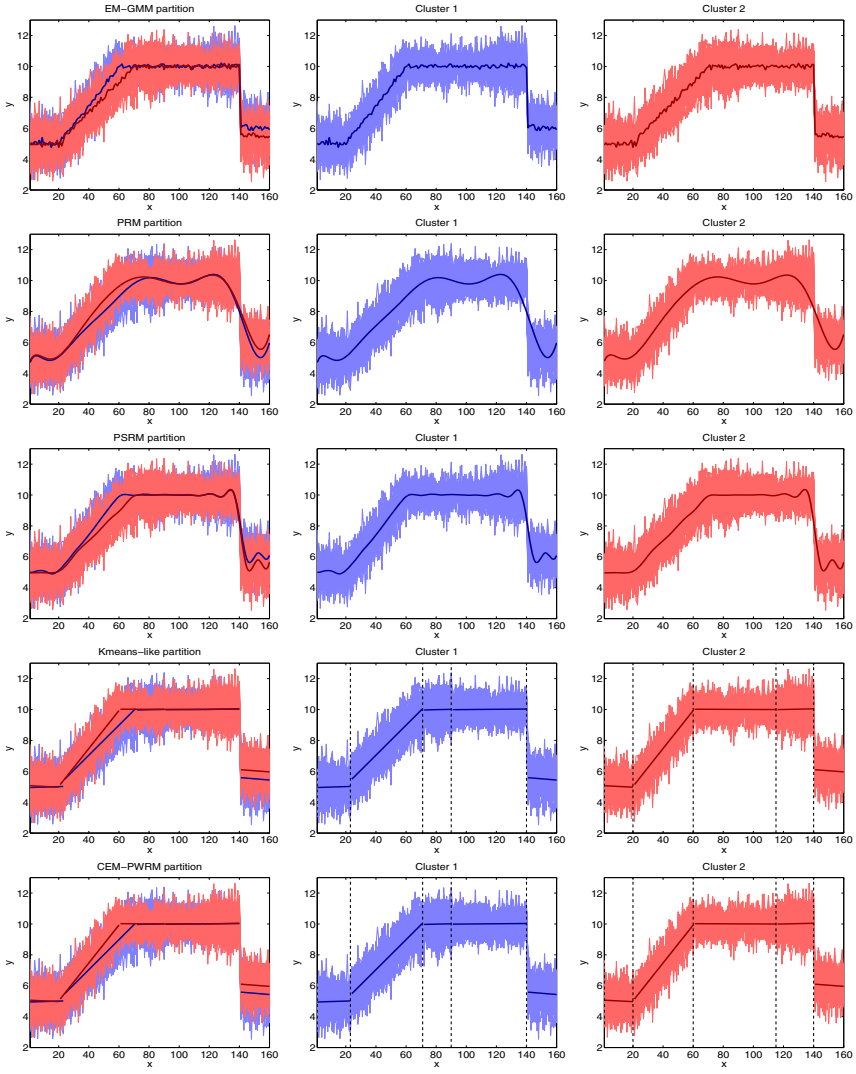

Figure 2. Clustering results and the corresponding cluster prototypes obtained with EM-GMM (spherical model), EM-PRM, EM-PSRM, and the corresponding cluster segmentations obtained with K-means-like and CEM-PWRM. For the color version of this figure, the reader is referred to the web version of this article.

$\sigma_{13} = 0.7$ and $\sigma_{14} = 0.6$). Simulated curves according to this situation are shown in Figure 4. The clustering results for this example are shown in Figure 5. The misclassification error for this situation is 7% for the K-means-like approaches, and 3% for the proposed PWRM approach. For the other

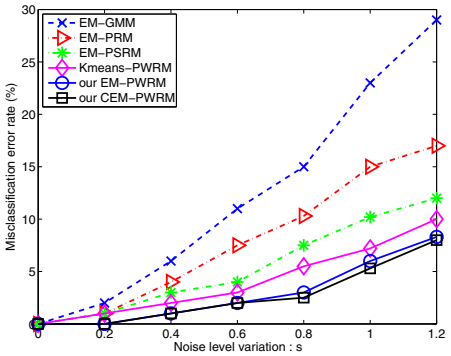

Figure 3. The misclassification error rate versus the noise level variation.

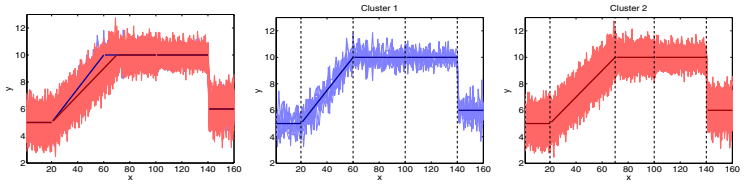

Figure 4. A two-class data set of simulated curves from a PWRM with non-uniform mixing proportions ($\alpha_1 = 0.2$, $\alpha_2 = 0.8$): the clusters colored according to the true partition, and the prototypes (left) and the true segmentation for cluster 1 (middle) and cluster 2 (right).For the color version of this figure, the reader is referred to the web version of this article.

approaches, the misclassification error is around 10% for both the PRM and the PSRM, while that for the GMM is of 20%. Another interesting point to see here is that the K-means based approach can fail in terms of segmentation. As can be seen in Figure 5 (top, right), the third and the fourth regime do not correspond to the actual ones (see Figure 4, middle). This is to be attributed to the fact that the K-means-like approach for PWRM is constrained in that it assumes the same proportion for each cluster, and does not sufficiently take into account the heteroskedasticity within each cluster, as does the proposed general probabilistic PWRM model.

6.2.4 Model Selection

In this section we give the results concerning the selection of the best values of the triplet (K,R,p) by using the ICL criterion as presented in Section 5.3. The values of $(K_{max},R_{max},p_{max})$ (respectively $(K_{min},R_{min},p_{min})$) were $(4,6,3)$ (respectively $(1,1,0)$). We note that for the K-means-like algorithm, the complete-data log-likelihood is $\mathscr{L}_c = -\frac{1}{2}E$ up to a constant term

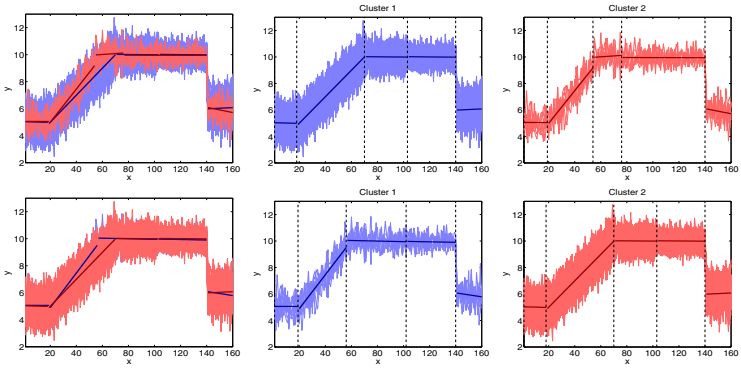

Figure 5. Results for the curves shown in Figure 4: Clustering results and the corresponding cluster prototypes and cluster segmentations obtained with K-means-like (top) and the proposed CEM-PWRM (down). For the color version of this figure, the reader is referred to the web version of this article.

(see Equation (33)), where E is the criterion minimized by this approach which is given by Equation (3). The ICL criterion for this approach is therefore computed as $\mathrm{ICL}(K,R,p) = -\frac{E}{2} - \frac{\nu_\Psi \log(n)}{2}$, where $\nu_\Psi = \sum_{k=1}^{K} R_k(p+2) - K$ is the number of free parameters of the model and n is the sample size. The number of free model parameters in this case includes $\sum_{k=1}^{K} R_k(p+1)$ polynomial coefficients and $\sum_{k=1}^{K}(R_k - 1)$ transition points, the model being a constrained PWRM model (isotropic with identical mixing proportions).

For this experiment, we observed that the model with the highest percentage of selection corresponds to $(K,R,p) = (2,5,1)$ for the proposed EM-PWRM and CEM-PWRM approaches with respectively 81% and 85% of selection. While for the K-means-like approach, the same model $(K,R,p) = (2,5,1)$ has a percentage of selection of only 72%. The number of regimes is underestimated by only around 10% by the proposed approaches, while the number of clusters is correctly estimated. However, the K-means-like approach overestimates the number of clusters ($K = 3$) in 12% of the cases. These results illustrate an advantage of the fully probabilistic approach compared to that based on the K-means-like approach. We also note that the models with $K = 1,4$ and those with $R = 1,2$ were not selected (percentage of 0%) for all the models.

6.3 Application to Real Curves

In this section we apply the proposed approach to real curves issuing from three different data sets, and compare it to the alternatives. The studied curves are the railway switch curves, the Tecator curves and the Topex/consist satellite data as studied in Hébrail et al. (2010). The curves of

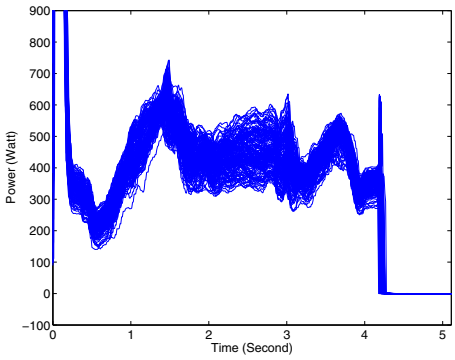

Figure 6. Railway switch curves.

these datasets are shown in Figure 6, Figure 8, and Figure 10. Note that the true partition for the three data sets is unknown.

6.3.1 Railway Switch Curves

The first studied curves are the railway switch curves issuing from a railway diagnosis application of the railway switch. Roughly, the railway switch is the component that enables (high speed) trains to be guided from one track to another at a railway junction, and is controlled by an electric motor. These curves are the signals of the consumed power during the switch operations. These curves present several changes in regime due to the successive mechanical motions involved in each switch operation (see Figure 6). The diagnosis task can be achieved through the analysis of these curves to identify possible faults. However, the large amount of data makes the manual labeling task onerous for the experts. Therefore, the main concern of this task is to propose a data preprocessing approach that allows automatically identifying homogeneous groups (without defects or with possible defects). The database used is composed of $n = 146$ real curves of $m = 511$ observations. We assume that in the database we have two clusters ($K = 2$). The first contains curves corresponding to an operating state without defect and the second contains curves corresponding to an operating state with a possible defect. The number of regression components was set to $R = 6$ in accordance with the number of electromechanical phases of a switch operation and the degree of the polynomial regression p was set to 3 which is appropriate for the different regimes in the curves. However, we note that no ground truth for this data set is available, neither regarding the classifications nor regarding the segmentation. This study could provide a preliminary result to help experts in labelling the data.

Table 3. Intra-cluster inertia for the switch curves.

EM-GMM	EM-PRM	EM-PSRM	K-means-like	CEM-PWRM
721.46	738.31	734.33	704.64	703.18

Figure 7 shows the graphical clustering results and the corresponding cluster prototypes for the real switch operation curves. We can see that the standard GMM clustering fails as it does not take into account the temporal aspect of the data, the obtained clusterings are not different and the mean curves are computed as over-all mean curves so that the obtained results are not very convincing. The results provided by the PRM and PSRM models are not convincing with regard to either the clustering or the approximation. However, the PWRM model clearly provides better results, since the cluster prototypes are more concordant with the real shape of the curves and, especially the proposed CEM-PWRM provides informative clusters. Indeed, it can be observed that for the CEM-PWRM approach, the curves of the first cluster (middle) and the second one (right) do not have the same characteristics since their shapes are clearly different. Therefore they may correspond to two different states of the switch mechanism. In particular, for the curves belonging to the first cluster (middle), it can be observed that something happened at around 4.2 seconds of the switch operation. According to the experts, this can be attributed to a fault in the measurement process, rather than a fault in the switch itself. The device used for measuring the power would have been used slightly differently for this set of curves. Since the true class labels are unknown, we consider the results of intra-class inertia, which are found to be more significant for these data than is the intra-class inertia of the extensions. The values of inertia corresponding to the results shown in Figure 7 are given in Table 3.The intra-class results confirm that the piecewise regression mixture model has an advantage at giving homogeneous and well approximated clusters from curves of regime changes.

6.3.2 Tecator Data

The Tecator data[1] consist of near infrared (NIR) absorbance spectra of 240 meat samples. The NIR spectra were recorded on a Tecator Infratec food and feed Analyzer working in the wavelength range $850 - 1050$ nm. The full Tecator data set contains $n = 240$ spectra with $m = 100$ for each spectrum, and is presented in Figure 8. This data set has been considered in Hébrail et al. (2010) and in our experiment we consider the same setting, that the data set is summarized with six clusters ($K = 6$), each cluster being composed of five linear regimes (segments) ($R = 5, p = 1$).

1. Tecator data are available at http://www.math.univ-toulouse.fr/staph/npfda/npfda-datasets.html.

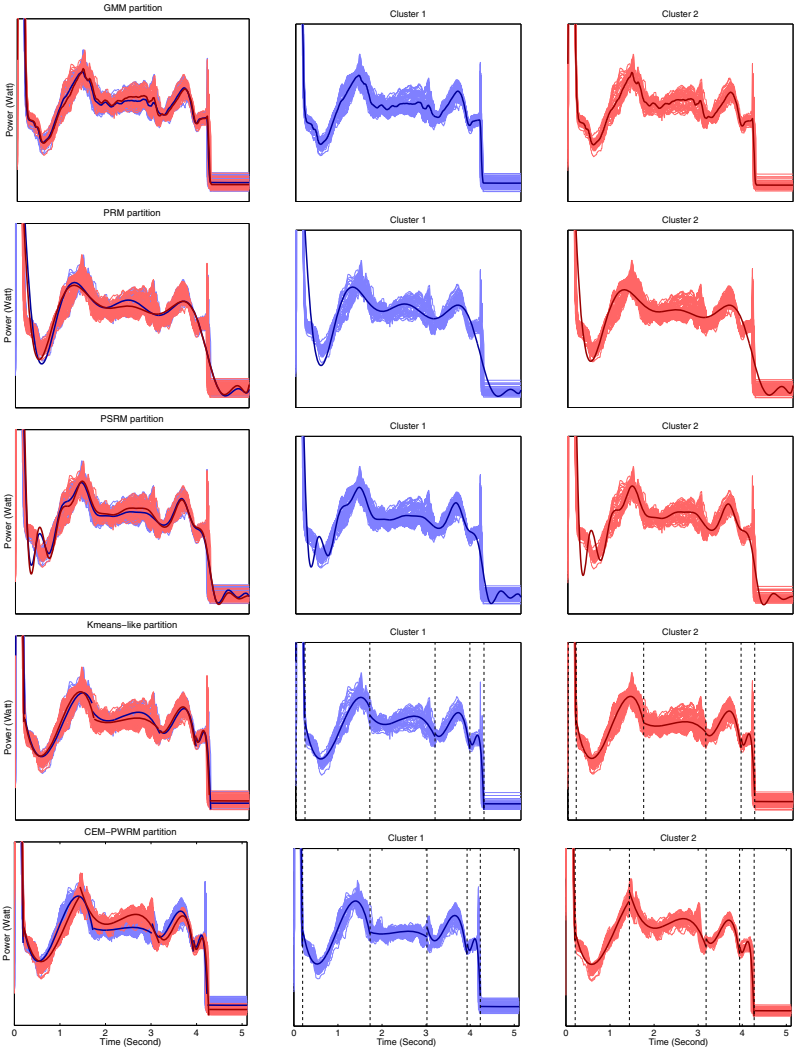

Figure 7. Clustering results and the corresponding cluster prototypes obtained with EM-GMM, EM-PRM, EM-PSRM, and the corresponding cluster segmentations obtained with K-means-like and CEM-PWRM. For the color version of this figure, the reader is referred to the web version of this article.

Figure 9 shows the clustering and segmentation results obtained by the proposed CEM-PWRM algorithm. One can see that the retrieved clusters are informative in the sense that the shapes of the clusters are clearly different, and the piecewise approximation is in concordance with the shape of each cluster. On the other hand, it can also be observed that this result is

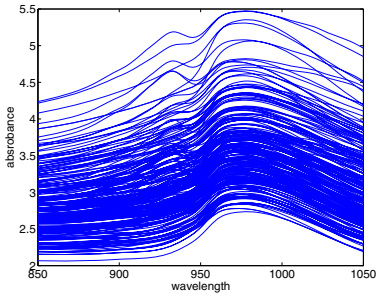

Figure 8. Tecator curves.

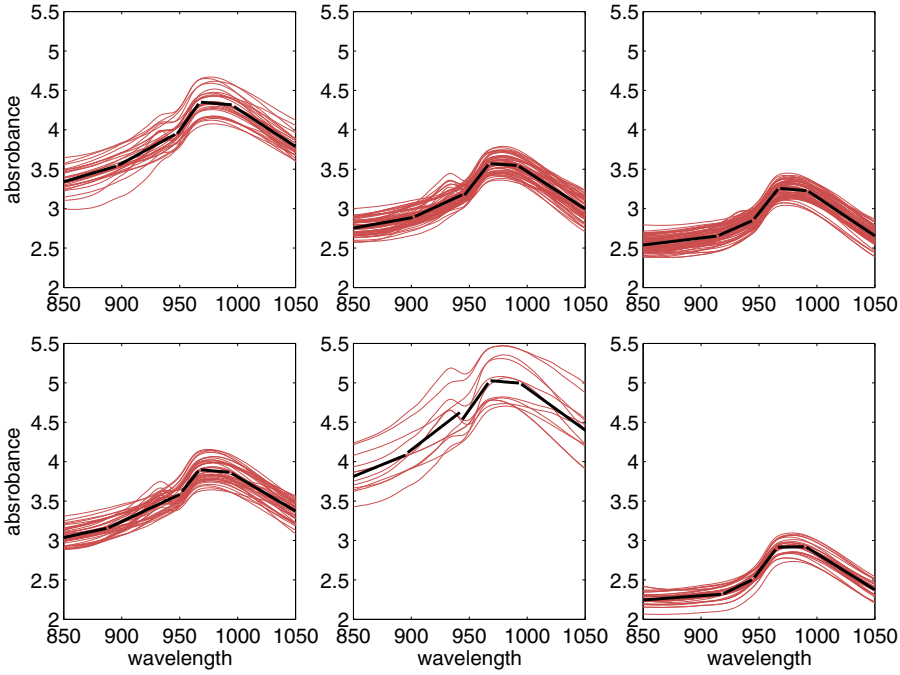

Figure 9. Clusters and the corresponding piecewise linear prototypes for each cluster obtained with the CEM-PWRM algorithm for the full Tecator data set.

very close to the one obtained by Hébrail et al. (2010) but using the K-means-like approach. This not surprising and confirms that our proposed CEM-PWRM algorithm is a probabilistic alternative for the K-means-like approach.

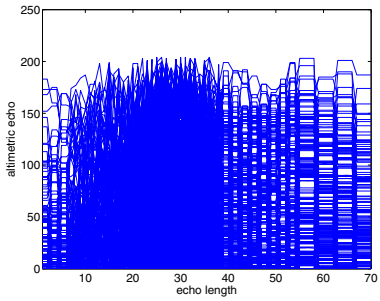

Figure 10. Topex/Poseidon satellite curves.

6.3.3 Topex/Poseidon Satellite Data

The Topex/Poseidon radar satellite data[2] were registered by the satellite Topex/Poseidon around an area of 25 kilometers upon the Amazon River. The data contain $n = 472$ waveforms of the measured echoes, sampled at $m = 70$ (number of echoes). The curves of this data set are shown in Figure 10. We employed the same number of clusters (20) and a piecewise linear approximation of four segments per cluster, as used in Hébrail et al. (2010). We note that, in our approach, we directly apply the proposed CEM-PWRM algorithm to the raw satellite data without a preprocessing step. However, in Hébrail et al. (2010), the authors used a two-fold scheme. They first performed a topographic clustering step using the Self Organizing Map (SOM), and then applied their K-means-like approach to the results of the SOM.

Figure 11 shows the clustering and segmentation results obtained with the proposed CEM-PWRM algorithm for the satellite data set. First, it can be observed that the provided clusters are clearly informative and reflect the general behavior of the hidden structure of this data set. The structure is indeed clearer with the mean curves of the clusters (prototypes) than with the raw curves. The piecewise approximation thus helps to better understand the structure of each cluster of curves from the obtained partition, and to more easily infer the general behavior of the data set. On the other hand, one can also see that this result is similar to the one found in Hébrail et al. (2010): most of the profiles are present in the two results. The slight difference can be attributed to the fact that the result in Hébrail et al. (2010) is provided from a two-stage scheme which includes an additional pre-clustering step using the SOM, rather than by directly applying the piecewise regression model to the raw data.

2. Satellite data are available at http://www.lsp.ups-tlse.fr/staph/npfda/npfda-datasets.html.

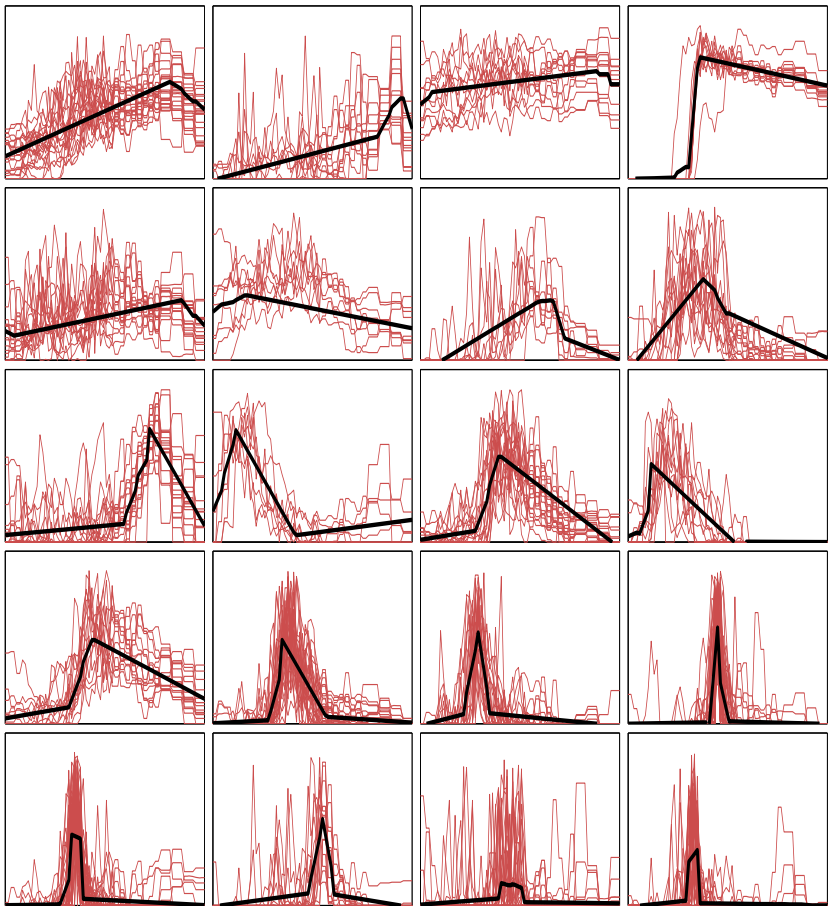

Figure 11. Clusters and the corresponding piecewise linear prototypes for each cluster obtained with the CEM-PWRM algorithm for the satellite data set.

7. Conclusions and Discussion

In this paper, we introduced a new probabilistic approach for simultaneous clustering and optimal segmentation of curves with regime changes. The proposed approach is a piecewise polynomial regression mixture (PWRM). We provided two algorithms to estimate the parameters of the model. The first (EM-PWRM) consists in using the EM algorithm to maximize the observed data log-likelihood and the second (CEM-PWRM)

is a CEM algorithm to maximize the complete-data log-likelihood. We showed that the CEM-PWRM algorithm is a general probabilistic-based version of the K-means-like algorithm of Hébrail et al. (2010). We conducted experiments on both simulated curves and real data sets to evaluate the proposed approach and compare it to alternatives, including the regression mixture, the spline regression mixtures and the standard GMM for multivariate data. The obtained results demonstrated the benefits of the proposed approach in terms of both curve clustering and piecewise approximation of the regimes of each cluster. In particular, the comparisons with the K-means-like algorithm approach confirm that the proposed CEM-PWRM is a general probabilistic alternative. In the experiments, the EM and CEM versions, in this clustering and segmentation context, provided similar results. It is worth mentioning that if the aim is primarily the density estimation, the EM version would be suggested since the CEM is known to provide inconsistent parameter estimates, the parameters being updated from only a subset of the data. However, CEM is known to be well-tailored to the purpose of segmentation and clustering.

We note that in some practical situations involving continuous functions, the proposed piecewise regression mixture, in its current formulation, may lead to discontinuities between the segments for the piecewise approximation. This can be easily avoided by slightly modifying the algorithm by adding an interpolation step, as performed in Hébrail et al. (2010). We also note that in this paper, we are interested in piecewise regimes which dot not overlap; only the clusters can overlap. However, one way to address regime overlap is to augment the number of regimes in the proposed approach so that a regime that overlaps (for example it occurs in two different time ranges) can be treated as two regimes. These two reconstructed non-overlapping regimes would have very close characteristics so as to correspond to a single overlapping regime.

References

ANDREWS, J., and MCNICHOLAS, P. (2014), "Variable Selection for Clustering and Classification", *Journal of Classification, 31(2)*, 136–153.

BANFIELD, J.D., and RAFTERY A.E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics, 49(3)*, 803–821.

BELLMAN, R. (1961), "On the Approximation of Curves by Line Segments Using Dynamic Programming", *Communications of the Association for Computing Machinery, 4(6)*, 284.

BIERNACKI, C., CELEUX, G., and GOVAERT, G. (2000), "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood", *IEEE PAMI, 22(7)*, 719–725.

BIERNACKI, C., CELEUX, G., and GOVAERT, G. (2003), "Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models", *Computational Statistics and Data Analysis, 41,* 561–575.

BOUVEYRON, C. (2014), "Adaptive Mixture Discriminant Analysis for Supervised Learning with Unobserved Classes", *Journal of Classification, 31(1)*, 49–84.

BOUVEYRON, C., and BRUNET, C. (2014), "Model-Based Clustering of High-Dimensional Data: A Review", *Computational Statistics & Data Analysis, 71*, 52–78.

BRAILOVSKY, V.L., and KEMPNER, Y. (1992), "Application of Piecewise Regression to Detecting Internal Structure of Signal", *Pattern Recognition, 25(11)*, 1361–1370.

CELEUX, G., and GOVAERT, G. (1992), "A Classification EM Algorithm for Clustering and Two Stochastic Versions", *Computational Statistics and Data Analysis, 14*, 315–332.

CELEUX, G., and GOVAERT, G. (1993), "Comparison of the Mixture and the Classification Maximum Likelihood in Cluster Analysis", *Journal of Statistical Computation and Simulation, 47*, 127–146.

CELEUX, G., and GOVAERT, G. (1995), "Gaussian Parsimonious Clustering Models", *Pattern Recognition, 28(5)*, 781–793.

CHAMROUKHI, F. (2010), "Hidden Process Regression for Curve Modeling, Classification and Tracking", Ph.D. thesis, Université de Technologie de Compiègne, France.

CHAMROUKHI, F., SAMÉ, A., GOVAERT, G., and AKNIN, P. (2009a), "A Regression Model with a Hidden Logistic Process for Feature Extraction from Time Series", *Neural Networks, 22(5-6)*, 593–602.

CHAMROUKHI, F., SAMÉ, A., GOVAERT, G., and AKNIN, P. (2009b), "Time Series Modeling by a Regression Approach Based on a Latent Process", *Neural Networks, 22(5-6)*, 593–602.

CHAMROUKHI, F., SAMÉ, A., GOVAERT, G., and AKNIN, P. (2010), "A Hidden Process Regression Model For Functional Data Description. Application to Curve Discrimination", *Neurocomputing, 73(7-9)*, 1210–1221.

CHAMROUKHI, F., SAMÉ, A., AKNIN, P., and GOVAERT, G. (2011), "Model-Based Clustering with Hidden Markov Model Regression for Time Series with Regime Changes", in *International Joint Conference on Neural Networks*, pp. 2814–2821.

CHAMROUKHI, F., HERVÉ, G., and SAMÉ, A. (2013), "Model-Based Functional Mixture Discriminant Analysis with Hidden Process Regression for Curve Classification", *Neurocomputing, 112*, 153–163.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Series B, 39(1)*, 1–38.

FEARNHEAD, P. (2006), "Exact and Efficient Bayesian Inference for Multiple Changepoint Problems", *Statistics and Computing, 16*, 203–213.

FEARNHEAD, P., and LIU, Z. (2007), "Online Inference for Multiple Changepoint Problems", *Journal of the Royal Statistical Society, Series B, 69*, 589–605.

FERRARI-TRECATE, G., and MUSELLI, M. (2002), "A New Learning Method for Piecewise Linear Regression", in *International Conference on Artificial Neural Networks*, pp. 28–30.

FRALEY, C., and RAFTERY, A.E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation", *Journal of the American Statistical Association, 97*, 611–631.

GAFFNEY, S., and SMYTH, P. (1999), "Trajectory Clustering with Mixtures of Regression Models", in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 63–72.

GAFFNEY, S.J. (2004), “Probabilistic Curve-Aligned Clustering and Prediction with Regression Mixture Models", PhD thesis, University of California, Irvine.

GAFFNEY, S.J., and SMYTH, P. (2004), “Joint Probabilistic Curve Clustering and Alignment", in Advances in Neural Information Processing Systems 17.

GANESALINGAM, S., and MCLACHLAN, G.J. (1978), “The Efficiency of a Linear Discriminant Function Based on Unclassified Initial Samples", *Biometrika, 65*, 658–662.

GANESALINGAM, S., and MCLACHLAN, G.J. (1979), “A Case Study of Two Clustering Methods Based on Maximum Likelihood", *Statistica Neerlandica, 33*, 81–90.

GOVAERT, G., INGRASSIA, S., and MCLACHLAN, G. (eds) (2015), “Special Issue on ‘New Trends on Model-Based Clustering and Classification‘ ", *Advances in Data Analysis and Classification, 9(4)*, 367–369.

GUI, J., and LI, H. (2003), “Mixture Functional Discriminant Analysis for Gene Function Classification Based on Time Course Gene Expression Data", in *Proceedings of the Joint Statistical Meeting (Biometric Section)*.

HÉBRAIL, G., HUGUENEY, B., LECHEVALLIER, Y., and ROSSI, F. (2010), “Exploratory Analysis of Functional Data via Clustering and Optimal Segmentation", *Neurocomputing 73(7–9)*, 1125–1141.

HUGUENEY, B., HÉBRAIL, G., LECHEVALLIER, Y., and ROSSI, F. (2009), "Simultaneous Clustering and Segmentation for Functional Data", in *European Symposium on Artificial Neural Networks*, pp. 281–286.

INGRASSIA, S., MINOTTI, S., and VITTADINI, G. (2012), “Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions", *Journal of Classification, 29(3)*, 363–401.

INGRASSIA, S., PUNZO, A., VITTADINI, G., and MINOTTI, S. (2015), “The Generalized Linear Mixed Cluster-Weighted Model", *Journal of Classification, 32(1)*, 85–113.

JACQUES, J., and PREDA, C. (2014), “Model-Based Clustering for Multivariate Functional Data", *Computational Statistics & Data Analysis, 71*, 92–106.

JAMES, G.M., and SUGAR, C. (2003), “Clustering for Sparsely Sampled Functional Data", *Journal of the American Statistical Association, 98(462)*, 397–408.

LEE, S., and MCLACHLAN, G. (2014), “Finite Mixtures of Multivariate Skew *t*-Distributions: Some Recent and New Results", *Statistics and Computing. 24(2)*, 181–202.

LEE, S.X., and MCLACHLAN, G.J. (2013), “Model-Based Clustering and Classification with Non-Normal Mixture Distributions", *Statistical Methods and Applications, 22(4)*, 427–454.

LEE, S.X., and MCLACHLAN, G.J. (2015), “Finite Mixtures of Canonical Fundamental Skew *t*-Distributions", *Statistics and Computing, 24(2)*, 181–202.

LIU, X., and YANG, M. (2009), “Simultaneous Curve Registration and Clustering for Functional Data", *Computational Statistics and Data Analysis, 53(4)*, 1361–1376.

MCGEE, V.E., and CARLETON, W.T. (1970), “Piecewise Regression", *Journal of the American Statistical Association, 65*, 1109–1124.

MCLACHLAN, G., and BASFORD, K. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

MCLACHLAN, G.J. (1982), “The Classification and Mixture Maximum Likelihood Approaches to Cluster Analysis", in *Handbook of Statistics, Vol. 2*, eds. P. Krishnaiah and L. Kanal, pp. 199–208.

MCLACHLAN, G.J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

MCLACHLAN, G.J., and KRISHNAN, T. (2008), *The EM Algorithm and Extensions* (2nd ed.), New York: Wiley.

MCLACHLAN, G.J., and PEEL, D. (2000), *Finite Mixture Models*, New York: Wiley.

MELNYKOV, V. (2016), "Model-Based Biclustering of Clickstream Data", *Computational Statistics & Data Analysis, 93(C)*, 31–45.

MELNYKOV, V., and MAITRA, R. (2010), "Finite Mixture Models and Model-Based Clustering", *Statistics Surveys 4*, 80–116.

MURRAY, P.M., BROWNE, R.P., and MCNICHOLAS, P.D. (2014), "Mixtures of Skew-Factor Analyzers", *Computational Statistics & Data Analysis, 77*, 326–335.

NGUYEN, H.D., MCLACHLAN, G.J., and WOOD, I.A. (2016), "Mixtures of Spatial Spline Regressions for Clustering and Classification", *Computational Statistics and Data Analysis, 93*, 76–85.

PICARD, F., ROBIN, S., LEBARBIER, E., and DAUDIN, J.J. (2007) "A Segmentation/Clustering Model for the Analysis of Array CGH Data", *Biometrics, 63(3)*, 758–766.

RAMSAY, J.O., and SILVERMAN, B.W. (2005), *Functional Data Analysis*, Berlin: Springer.

SAMÉ, A., CHAMROUKHI, F., GOVAERT, G., and AKNIN, P. (2011) "Model-Based Clustering and Segmentation of Time Series with Changes in Regime", *Advances in Data Analysis and Classification, 5(4)*, 301–321.

SCHWARZ, G. (1978), "Estimating the Dimension of a Model", *Annals of Statistics, 6*, 461–464.

SCOTT, A.J., and SYMONS, M.J. (1971), "Clustering Methods Based on Likelihood Ratio Criteria", *Biometrics, 27*, 387–397.

SHI, J.Q., and WANG, B. (2008), "Curve Prediction and Clustering with Mixtures of Gaussian Process Functional Regression Models", *Statistics and Computing, 18(3)*, 267–283.

SMYTH, P. (1996). "Clustering Sequences with Hidden Markov Models", in *Advances in Neural Information Processing Systems 9, NIPS,* pp. 648–654.

STEINLEY, D., and BRUSCO M.J. (2007), "Initializing *k*-Means Batch Clustering: A Critical Evaluation of Several Techniques", *Journal of Classification, 24*, 99–121.

STONE, H. (1961), "Approximation of Curves by Line Segments", *Mathematics of Computation, 15(73)*, 40–47.

TANG, Y., BROWNE, R.P., and MCNICHOLAS, P.D. (2015), "Model Based Clustering of High-Dimensional Binary Data", *Computational Statistics & Data Analysis, 87*, 84–101.

TITTERINGTON, D., SMITH, A., and MAKOV, U. (1985.) *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley & Sons.

WOLFE, J.H. (1970), "Pattern Clustering by Multivariate Mixture Analysis", *Multivariate Behavior Research, 5*, 329–359.

XIONG, Y., and YEUNG, D.Y. (2004), "Time Series Clustering with ARMA Mixtures", *Pattern Recognition, 37(8)*, 1675–1689.