Time series modeling by a regression approach based on a latent process

Faicel Chamroukhi[*,a,b], Allou Samé[a], Gérard Govaert[b], Patrice Aknin[a]

[a]*French National Institute for Transport and Safety Research (INRETS)*
Laboratory of New Technologies (LTN)
2 Rue de la Butte Verte,
93166 Noisy-le-Grand Cedex (France)
[b]*Compiègne University of Technology*
HEUDIASYC Laboratory, UMR CNRS 6599
BP 20529, 60205 Compiègne Cedex (France)

Abstract

Time series are used in many domains including finance, engineering, economics and bioinformatics generally to represent the change of a measurement over time. Modeling techniques may then be used to give a synthetic representation of such data. A new approach for time series modeling is proposed in this paper. It consists of a regression model incorporating a discrete hidden logistic process allowing for activating smoothly or abruptly different polynomial regression models. The model parameters are estimated by the maximum likelihood method performed by a dedicated Expectation Maximization (EM) algorithm. The M step of the EM algorithm uses a multi-class Iterative Reweighted Least-Squares (IRLS) algorithm to estimate the hidden process parameters. To evaluate the proposed approach, an experimental study on simulated data and real world data was performed using two alternative approaches: a heteroskedastic piecewise regression model using a global optimization algorithm based on dynamic programming, and a

[*]Corresponding author:
Faicel Chamroukhi
INRETS, 2 Rue de la Butte Verte,
93166 Noisy-le-Grand Cedex, France
Tel: +33(1) 45 92 56 46
Fax: +33(1) 45 92 55 01
Email address: `faicel.chamroukhi@inrets.fr` (Faicel Chamroukhi)

Hidden Markov Regression Model whose parameters are estimated by the Baum-Welch algorithm. Finally, in the context of the remote monitoring of components of the French railway infrastructure, and more particularly the switch mechanism, the proposed approach has been applied to modeling and classifying time series representing the condition measurements acquired during switch operations.

1. Introduction

Time series occur in many domains including finance, engineering, economics, bioinformatics, and they generally represent the change of a measurement over time. Modeling techniques may then be used to give a synthetic representation of such data. This work relates to the diagnosis of the French railway switches (or points) which enable trains to be guided from one track to another at a railway junction. For this purpose, condition measurements acquired during switch operations are classified into predefined classes. Each measurement represents the electrical power consumed during a switch operation (see Fig. 1).
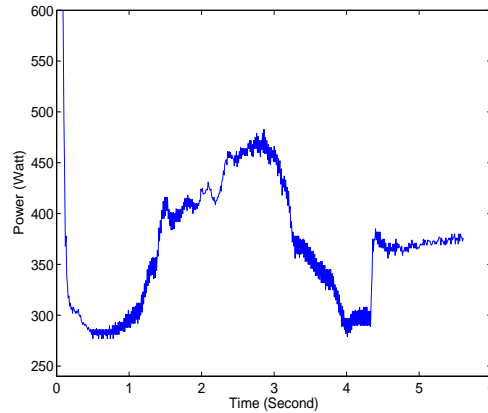

Figure 1: A signal showing the electrical power consumed during a switch operation.

The diagnosis task was performed by means of a two-step process: feature extraction from the switch operation signals and the implementation of a supervised learning algorithm to learn the parameters of the operating

classes of the switch mechanism. In this paper we propose a new method for modeling switch operation signals.

Switch operations signals can be seen as time series presenting non-linearities and various changes in regime. In a context of this type, basic parametric methods based on linear or polynomial regression are not adapted. A piecewise regression model may be used as an alternative (McGee and Carleton, 1970; Brailovsky and Kempner, 1992; Ferrari-Trecate and Muselli, 2002). Piecewise polynomial regression is a parametrization and segmentation method that partitions the data into K segments, each segment being characterized by its mean polynomial curve and its variance. For this type of modeling, the parameters estimation can be exactly performed using dynamic programming algorithm (Bellman, 1961) such as Fisher's algorithm (Fisher, 1958). This algorithm optimizes an additive cost function over all the segments of the time series (Lechevalier, 1990; Brailovsky and Kempner, 1992). However, it is well-known that dynamic programming procedures are computationally expensive. An iterative algorithm can be derived to improve the running time of Fisher's algorithm as proposed by Samé et al. (2007). This approach iteratively estimates the regression model parameters and the partition of the time series. The standard piecewise regression model usually assumes that noise variance is uniform in all the segments (homoskedastic model) (Brailovsky and Kempner, 1992; Ferrari-Trecate and Muselli, 2002; Ferrari-Trecate et al., 2002; Samé et al., 2007). However, in this paper we shall consider a heteroskedastic piecewise polynomial regression model. Another alternative approach is to use a Hidden Markov Regression Model (Fridman, 1993) whose parameters are estimated by the Baum-Welch algorithm (Baum et al., 1970). However the piecewise and Hidden Markov Regression approaches are more adapted for modeling time series presenting abrupt changes and may be less efficient for time series including regimes with smooth transitions.

The method we propose for time series modeling is based on a specific regression model incorporating a discrete hidden process allowing for abrupt or smooth transitions between different regression models. This approach is related to the switching regression model introduced by Quandt and Ramsey (1978) and is very linked to the Mixture of Experts (ME) model developed by Jordan and Jacobs (1994) by the using of a time-dependent logistic transition function. The ME model, as discussed in (Waterhouse, 1997), uses a conditional mixture modeling where the model parameters are estimated by the Expectation Maximization (EM) algorithm (Dempster and Rubin,

3

1977; McLachlan and Krishnan, 1997). Once the model parameters of the proposed regression model with hidden process are estimated, they are used as the feature vector for each signal. The parameters of the different operating classes (no defect, minor defect and critical defect) are then learnt from a labelled collection of signals using Mixture Discriminant Analysis (MDA) (Hastie and Tibshirani, 1996). Based on the operating classes parameters, a new signal is classified by using the Maximum A Posteriori (MAP) rule. The good performance of the proposed approach has been demonstrated by an experimental study carried out on real measured signals covering a wide range of defects.

This paper is organized as follows. Section 2 provides an account of the heteroskedastic piecewise polynomial regression model, and the parameter estimation technique this uses based on a dynamic programming procedure. Section 3 presents the Hidden Markov Regression Model whose parameters are estimated by the Expectation Maximization Baum-Welch algorithm. Section 4 introduces the proposed model and describes parameters estimation by means of the EM algorithm. Section 5 deals with the experimental study that assesses the performance of the proposed approach in terms of signal modeling and section 6 describes the application of the proposed technique to switch operation signals modeling and classification.

2. The piecewise polynomial regression model

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be n real observations of a signal or a time series where x_i is observed at time t_i. The piecewise polynomial regression model assumes that the time series incorporates K polynomial regimes on K intervals whose bounds indexes can be denoted by $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K+1})$ with $\gamma_1 = 0$ and $\gamma_{K+1} = n$. This defines a partition of the time series into K polynomial segments $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)$ of lengths $n_1, \ldots, n_K$ where $\boldsymbol{x}_k = \{x_i | i \in I_k\}$ is the set of elements in segment k whose indexes are $I_k = (\gamma_k, \gamma_{k+1}]$.

Standard polynomial regression models are homoskedastic models as they assume that the different polynomial regression models have the same noise variance. In our case we shall consider the more general framework of a heteroskedastic model which allows the noise level to vary between the different polynomial regression models. It can be defined as follows:

$$\forall i = 1, \ldots, n, \quad x_i = \boldsymbol{\beta}_k^T \boldsymbol{r}_i + \sigma_k \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, 1), \tag{1}$$

4

where k satisfies $i \in I_k$, $\boldsymbol{\beta}_k$ is the $(p+1)$-dimensional coefficients vector of a p degree polynomial associated with the kth segment with $k \in \{1, \ldots, K\}$, $\boldsymbol{r}_i = (1, t_i, t_i^2 \ldots, t_i^p)^T$ is the time dependent $(p+1)$-dimensional covariate vector associated to the parameter $\boldsymbol{\beta}_k$ and the ε_i are independent random variables with a standard Gaussian distribution representing the additive noise in each segment k.

2.1. Maximum likelihood estimation for the piecewise polynomial regression model

With this model, the parameters can be denoted by $(\boldsymbol{\psi}, \boldsymbol{\gamma})$ where $\boldsymbol{\psi} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2)$ is the set of polynomial coefficients and noise variances, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K+1})$ is the set of transition points. Parameter estimation is performed by maximum likelihood. We assume a conditional independence of the data. Thus, according to the model defined by equation (1), it can be proved that within each segment k, x_i has a Gaussian distribution with mean $\boldsymbol{\beta}_k^T \boldsymbol{r}_i$ and variance σ_k^2, and therefore, the log-likelihood of the parameter vector $(\boldsymbol{\psi}, \boldsymbol{\gamma})$ characterizing the piecewise regression model is the sum of the local log-likelihoods over the K segments that can be written as follows

$$
\begin{aligned}
L(\boldsymbol{\psi}, \boldsymbol{\gamma}; \mathbf{x}) &= \log p(\mathbf{x}; \boldsymbol{\psi}, \boldsymbol{\gamma}) \\
&= \sum_{k=1}^{K} \sum_{i \in I_k} \log \mathcal{N} \left(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2 \right).
\end{aligned} \tag{2}
$$

Maximizing this log-likelihood is equivalent to minimizing with respect to $\boldsymbol{\psi}$ and $\boldsymbol{\gamma}$ the criterion

$$
J(\boldsymbol{\psi}, \boldsymbol{\gamma}) = \sum_{k=1}^{K} \left[\frac{1}{\sigma_k^2} \sum_{i \in I_k} \left(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i \right)^2 + n_k \log \sigma_k^2 \right], \tag{3}
$$

where n_k is the number of elements in segment k.

Since the criterion J is additive over the K segments, the Fisher's algorithm (Fisher, 1958; Lechevalier, 1990), which consists in a dynamic programming procedure (Bellman, 1961; Brailovsky and Kempner, 1992), can be used to perform the global minimization. This dynamical procedure has a time complexity of $O(Kp^2n^2)$ which can be computationally expensive for large sample sizes.

5

2.2. Time series approximation and segmentation with the piecewise regression model

Once the parameters have been estimated, a segmentation of the time series, equivalently represented by the classes vector $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_n)$, where $\hat{z}_i \in \{1, \ldots, K\}$, can be derived by setting $\hat{z}_i = k$ if $i \in (\hat{\gamma}_k; \hat{\gamma}_{k+1}]$, the parameters $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}})$ being the parameters provided by the dynamic programming procedure.

An approximation of the time series is then given by $\hat{x}_i = \sum_{k=1}^{K} \hat{z}_{ik} \hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i$, where $\hat{z}_{ik} = 1$ if $\hat{z}_i = 1$ and $\hat{z}_{ik} = 0$ otherwise. The vectorial formulation of the approximated time series $\hat{\mathbf{x}}$ can be written as:

$$\hat{\mathbf{x}} = \sum_{k=1}^{K} \hat{Z}_k \mathbf{T} \hat{\boldsymbol{\beta}}_k, \tag{4}$$

where $\hat{Z}_k$ is a diagonal matrix whose diagonal elements are $(\hat{z}_{1k}, \ldots, \hat{z}_{nk})$, and

$$\mathbf{T} = \begin{bmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^p \\ 1 & t_2 & t_2^2 & \ldots & t_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 & \ldots & t_n^p \end{bmatrix}$$

is the $[n \times (p+1)]$ regression matrix.

3. The Hidden Markov Regression Model

This section recalls the Hidden Markov Regression Model (HMRM) (Fridman, 1993). Owing to the fact that the real signals we want to model consist of successive phases, order constraints are assumed for the hidden states in the HMRM.

3.1. A general description of Hidden Markov Regression Models

In a Hidden Markov Regression Model, the time series is represented as a sequence of observed variables $\mathbf{x} = (x_1, \ldots, x_n)$, where x_i is observed at time t_i and assumed to be generated by the following regression model (Fridman, 1993):

$$\forall i = 1, \ldots, n, \quad x_i = \boldsymbol{\beta}_{z_i}^T \boldsymbol{r}_i + \sigma_{z_i} \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, 1), \tag{5}$$

where z_i is a discrete hidden variable taking its values in the set $\{1, \ldots, K\}$.

6

The HMRM assumes that the hidden variable $\mathbf{z} = (z_1, \ldots, z_n)$ is a homogeneous Markov chain where the variable z_i controls the switching from one polynomial regression model to another of K models at each time t_i. The distribution of the latent sequence $\mathbf{z} = (z_1, \ldots, z_n)$ is defined as:

$$
\begin{aligned}
p(\mathbf{z}; \pi, A) &= p(z_1; \pi) \prod_{i=2}^{n} p(z_i|z_{i-1}; A) \\
&= \prod_{k=1}^{K} \pi_k^{z_{1k}} \prod_{i=2}^{n} \prod_{k=1}^{K} \left[\prod_{\ell=1}^{K} A_{\ell k}^{z_{(i-1)\ell}} \right]^{z_{ik}},
\end{aligned}
\tag{6}
$$

where

- $\pi = (\pi_1, \ldots, \pi_K)$ is the initial distribution of z_i, with $\pi_k = p(z_1 = k)$ for $k \in \{1, \ldots, K\}$;

- $A = (A_{\ell k})_{1 \leq \ell, k \leq K}$ where $A_{\ell k} = p(z_i = k | z_{i-1} = \ell)$ is the matrix of transition probabilities;

- $z_{ik} = 1$ if $z_i = k$ (i.e if x_i is generated by the kth regression model) and $z_{ik} = 0$ otherwise.

3.2. Parameter estimation of the Hidden Markov Regression Model

From model defined by equation (5), it can be proved that, conditionally on a regression model k ($z_i = k$), x_i has a Gaussian distribution with mean $\boldsymbol{\beta}_k^T \boldsymbol{r}_i$ and variance σ_k^2. Thus, the HMRM is parameterized by the parameter vector $\boldsymbol{\Psi} = (\pi, A, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2)$. The parameter vector $\boldsymbol{\Psi}$ is estimated by the maximum likelihood method. The log-likelihood to be maximized in this case is written as:

$$
\begin{aligned}
L(\boldsymbol{\Psi}; \mathbf{x}) &= \log p(\mathbf{x}; \boldsymbol{\Psi}) \\
&= \log \sum_{\mathbf{z}} p(z_1; \pi) \prod_{i=2}^{n} p(z_i|z_{i-1}; A) \prod_{i=1}^{n} \mathcal{N}(x_i; \boldsymbol{\beta}_{z_i}^T, \sigma_{z_i}^2).
\end{aligned}
\tag{7}
$$

Since the log-likelihood can not be maximized directly, this is done by the EM algorithm (Dempster and Rubin, 1977), which is known as the Baum-Welch algorithm (Baum et al., 1970) in the context of HMMs. It can easily be verified that, in a regression context, the Baum-Welch algorithm has a time complexity of $O(IKp^2n)$, where I is the number of iterations of the algorithm.

7

3.3. A HMRM with order constraints

Since the switch operation signals we aim to model consist of successive phases, we impose the following constraints on the transition probabilities:

$$p(z_i = k|z_{i-1} = \ell) = 0 \quad \text{if} \quad k < \ell , \tag{8}$$

and

$$p(z_i = k|z_{i-1} = \ell) = 0 \quad \text{if} \quad k > \ell + 1. \tag{9}$$

These constraints imply that no transitions are allowed for the phases whose indices are lower that the current phase (equation 8) and no jumps of more than one state are possible (equation 9). This constrained model is a particular case of the well known left-right model (Rabiner, 1989).

3.4. Time series approximation and segmentation with the HMRM

To approximate the time series, at each time t_i we combine the different regression models using the filtering probabilities denoted by ω_{ik} for the kth regression model. The filtering probability is the probability $\omega_{ik} = p(z_i = k|x_1, \ldots, x_i; \mathbf{\Psi})$ that x_i will be generated by the regression model k given the observations $(x_1, \ldots, x_i)$ that occur until time t_i. It can be computed using the so-called "forward" probabilities (Rabiner, 1989). Thus, the filtered time series $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$, which is common way to approximate the time series $\mathbf{x}$, is given by:

$$\hat{x}_i = \sum_{k=1}^{K} \hat{\omega}_{ik}\hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i ; \; i = 1, \ldots, n, \tag{10}$$

where $\hat{\mathbf{\Psi}} = (\hat{\pi}, \hat{A}, \hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2)$ and $\hat{\omega}_{ik}$ are respectively the parameter vector and the filtered probability obtained using the EM (Baum-Welch) algorithm. The vectorial formulation of the approximated time series $\hat{\mathbf{x}}$ can be written as:

$$\hat{\mathbf{x}} = \sum_{k=1}^{K} \hat{\mathcal{W}}_k \mathbf{T} \hat{\boldsymbol{\beta}}_k, \tag{11}$$

where $\hat{\mathcal{W}}_k$ is a diagonal matrix whose diagonal elements are $(\hat{\omega}_{1k}, \ldots, \hat{\omega}_{nk})$, and $\mathbf{T}$ is the $[n \times (p+1)]$ regression matrix. This approximation will be taken as the denoised signal.

8

On the other hand, a segmentation of the time series can be deduced by computing the label $\hat{z}_i$ of x_i using the Maximum A Posteriori (MAP) rule as follows:

$$\hat{z}_i = \arg \max_{1 \leq k \leq K} \hat{\tau}_{ik} \; ; \; \forall i = 1, \ldots, n, \tag{12}$$

where $\tau_{ik} = p(z_i = k | \mathbf{x}; \boldsymbol{\Psi})$ is the posterior probability that x_i originates from the kth regression model. Notice that τ_{ik} can be computed using the "forward" and "backward" probabilities (Rabiner, 1989).

4. The proposed regression model with a hidden logistic process

The proposed regression model introduced in this section is defined, as for the HMRM model, by equation (5), where a logistic process is used to model the hidden sequence $\mathbf{z} = (z_1, \ldots, z_n)$.

4.1. The hidden logistic process

This section defines the probability distribution of the process $\mathbf{z} = (z_1, \ldots, z_n)$ that allows the switching from one regression model to another.

The proposed hidden logistic process assumes that the variables z_i, given the vector $\boldsymbol{t} = (t_1, \ldots, t_n)$, are generated independently according to the multinomial distribution $\mathcal{M}(1, \pi_{i1}(\mathbf{w}), \ldots, \pi_{iK}(\mathbf{w}))$, where

$$\pi_{ik}(\mathbf{w}) = p(z_i = k; \mathbf{w}) = \frac{\exp\left(\boldsymbol{w}_k^T \boldsymbol{v}_i\right)}{\sum_{\ell=1}^{K} \exp\left(\boldsymbol{w}_\ell^T \boldsymbol{v}_i\right)}, \tag{13}$$

is the logistic transformation of a linear function of the time-dependent co-variate $\boldsymbol{v}_i = (1, t_i, t_i^2, \ldots, t_i^q)^T$, $\boldsymbol{w}_k = (\boldsymbol{w}_{k0}, \ldots, \boldsymbol{w}_{kq})^T$ is the $(q+1)$-dimensional coefficients vector associated with the covariate $\boldsymbol{v}_i$ and $\mathbf{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$. Thus, given the vector $\boldsymbol{t} = (t_1, \ldots, t_n)$, the distribution of $\mathbf{z}$ can be written as:

$$p(\mathbf{z}; \mathbf{w}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{\exp\left(\boldsymbol{w}_k^T \boldsymbol{v}_i\right)}{\sum_{\ell=1}^{K} \exp\left(\boldsymbol{w}_\ell^T \boldsymbol{v}_i\right)} \right)^{z_{ik}}, \tag{14}$$

where $z_{ik} = 1$ if $z_i = k$ i.e when x_i is generated by the kth regression model, and 0 otherwise.

The relevance of the logistic transformation in terms of flexibility of transition can be illustrated through simple examples with $K = 2$ components. In this case, only the probability $\pi_{i1}(\mathbf{w}) = \frac{exp(\boldsymbol{w}_1^T \boldsymbol{v}_i)}{1 + exp(\boldsymbol{w}_1^T \boldsymbol{v}_i)}$ should be described,

since $\pi_{i2}(\mathbf{w}) = 1 - \pi_{i1}(\mathbf{w})$. The first example is designed to show the effect of the dimension q of $\boldsymbol{w}_k$ on the temporal variation of the probabilities π_{ik}. We consider different values of the dimension q $(q = 0, 1, 2)$ of $\boldsymbol{w}_k$.

As shown in Fig. 2, the dimension q controls the number of temporal transitions of π_{ik}. In fact, the larger the dimension of $\boldsymbol{w}_k$, the more complex the temporal variation of π_{ik}. More particularly, if the goal is to segment the signals into contiguous segments, the dimension q of $\boldsymbol{w}_k$ must be set to 1.
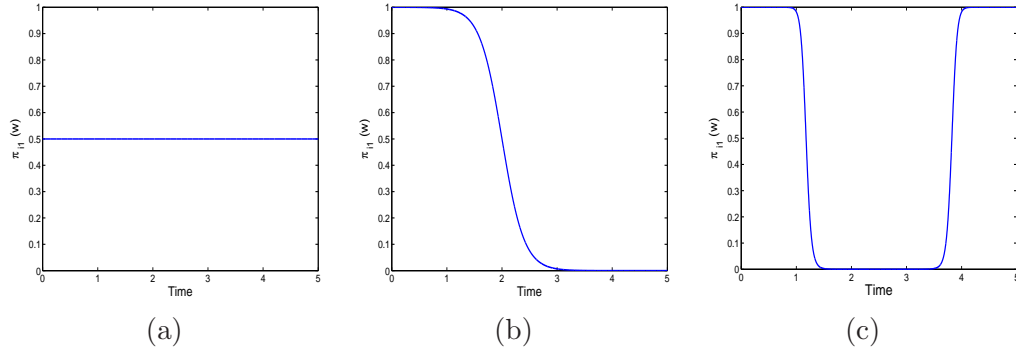
| (a) | (b) | (c) |

Figure 2: Variation of $\pi_{i1}(\mathbf{w})$ over time for different values of the dimension q of $\boldsymbol{w}_1$, for $K = 2$ and (a) $q = 0$ and $\boldsymbol{w}_1 = 0$, (b) $q = 1$ and $\boldsymbol{w}_1 = (10, -5)^T$ and (c) $q = 2$ and $\boldsymbol{w}_1 = (-10, -20, -4)^T$.

For a fixed dimension q of the parameter $\boldsymbol{w}_k$, the variation of the proportions $\pi_{ik}(\mathbf{w})$ over time, in relation to the parameter $\boldsymbol{w}_k$, is illustrated by an example of 2 classes with $q = 1$. For this purpose, we use the parametrization $\boldsymbol{w}_k = \lambda_k(\alpha_k, 1)^T$ of $\boldsymbol{w}_k$, where $\lambda_k = \boldsymbol{w}_{k1}$ and $\alpha_k = \frac{\boldsymbol{w}_{k0}}{\boldsymbol{w}_{k1}}$. As shown in Fig. 3 (a), the parameter λ_k controls the quality of transitions between classes, the higher absolute value of λ_k, the more abrupt the transition between the z_i, while the parameter α_k controls the transition time point via the inflexion point of the curve (see Fig. 3 (b)).

In this particular regression model, the variable z_i controls the switching from one regression model to another of K regression models at each time t_i. Therefore, unlike basic polynomial regression models, which assume uniform regression parameters over time, the proposed model permits the polynomial coefficients to vary over time by switching from one regression model to another.
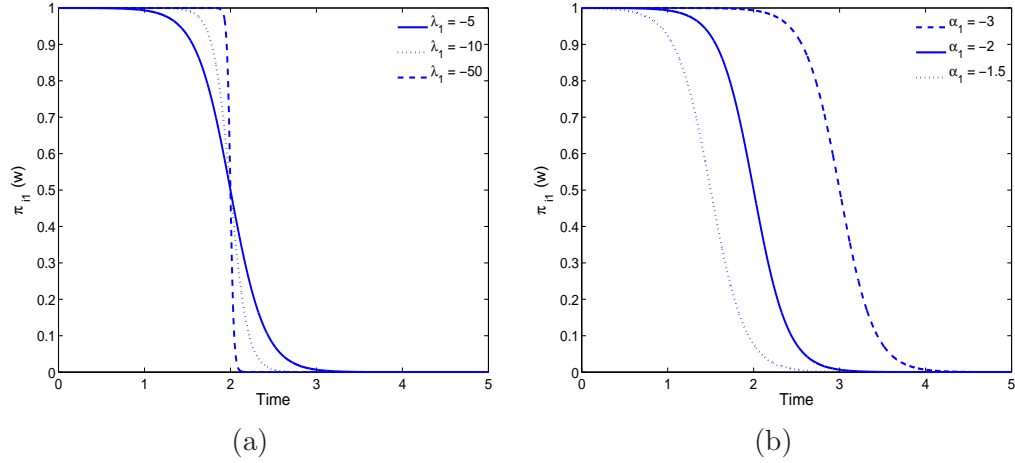
10

Figure 3: Variation of $\pi_{i1}(\mathbf{w})$ over time for a dimension $q = 1$ of $\boldsymbol{w}_1$ and (a) different values of $\lambda_1 = \boldsymbol{w}_{11}$ with $\alpha_1 = -2$ and (b) different values of $\alpha_1 = \frac{\boldsymbol{w}_{10}}{\boldsymbol{w}_{11}}$ with $\lambda_1 = -5$.

4.2. The generative model for signals

The generative model that produces a signal from a fixed parameter $\boldsymbol{\theta} = \{\boldsymbol{w}_k, \boldsymbol{\beta}_k, \sigma_k^2; k = 1, \ldots, K\}$ consists of 2 steps:

- generate the hidden process $\mathbf{z} = (z_1, \ldots, z_n)$ according to the multinomial distribution $z_i \sim \mathcal{M}(1, \pi_{i1}(\mathbf{w}), \ldots, \pi_{iK}(\mathbf{w}))$,

- generate each observation x_i according to the Gaussian distribution $\mathcal{N}(\cdot; \boldsymbol{\beta}_{z_i}^T \boldsymbol{r}_i, \sigma_{z_i}^2)$.

4.3. Parameter estimation

From the proposed model, it can be proved that, conditionally on a regression model k, x_i is distributed according to a normal density with mean $\boldsymbol{\beta}_k^T \boldsymbol{r}_i$ and variance σ_k^2. Thus, it can be proved that x_i is distributed according to the normal mixture density

$$p(x_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{w}) \mathcal{N}(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2), \tag{15}$$

where $\boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2)$ is the parameter vector to be estimated. The parameter $\boldsymbol{\theta}$ is estimated by the maximum likelihood method. As in the classic regression models we assume that, given $\boldsymbol{t} =$

11

$(t_1, \ldots, t_n)$, the ε_i are independent. This also implies the independence of x_i $(i = 1, \ldots, n)$. The log-likelihood of $\boldsymbol{\theta}$ is then written as:

$$
\begin{aligned}
L(\boldsymbol{\theta}; \mathbf{x}) &= \log \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_{ik}(\mathbf{w}) \mathcal{N}\left(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2\right). \quad (16)
\end{aligned}
$$

Since the direct maximization of this likelihood is not straightforward, it is maximized with the Expectation Maximization (EM) algorithm (Dempster and Rubin, 1977; McLachlan and Krishnan, 1997).

4.4. The dedicated EM algorithm

The proposed EM algorithm starts from an initial parameter $\boldsymbol{\theta}^{(0)}$ and alternates the two following steps until convergence:

4.4.1. E Step (Expectation)

This step consists in computing the expectation of the complete log-likelihood $\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$, given the observations and the current value $\boldsymbol{\theta}^{(m)}$ of the parameter $\boldsymbol{\theta}$ (m being the current iteration):

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) &= E\left[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) | \mathbf{x}; \boldsymbol{\theta}^{(m)}\right] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} E(z_{ik}|x_i; \boldsymbol{\theta}^{(m)}) \log\left[\pi_{ik}(\mathbf{w}) \mathcal{N}(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2)\right] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log\left[\pi_{ik}(\mathbf{w}) \mathcal{N}\left(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2\right)\right] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_{ik}(\mathbf{w}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \mathcal{N}\left(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2\right), (17)
\end{aligned}
$$

where

$$
\tau_{ik}^{(m)} = p(z_{ik} = 1|x_i; \boldsymbol{\theta}^{(m)}) = \frac{\pi_{ik}(\mathbf{w}^{(m)}) \mathcal{N}(x_i; \boldsymbol{\beta}_k^{T(m)} \boldsymbol{r}_i, \sigma_k^{2(m)})}{\sum_{\ell=1}^{K} \pi_{i\ell}(\mathbf{w}^{(m)}) \mathcal{N}(x_i; \boldsymbol{\beta}_\ell^{T(m)} \boldsymbol{r}_i, \sigma_\ell^{2(m)})} \quad (18)
$$

is the posterior probability that x_i originates from the kth regression model. As shown in the expression for Q, this step simply requires the computation of $\tau_{ik}^{(m)}$.

12

4.4.2. M step (Maximization)

In this step, the value of the parameter $\boldsymbol{\theta}$ is updated by computing the parameter $\boldsymbol{\theta}^{(m+1)}$ maximizing the conditional expectation Q with respect to $\boldsymbol{\theta}$. To perform this maximization, it can be observed that Q is written as:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = Q_1(\mathbf{w}) + \sum_{k=1}^{K} Q_2(\boldsymbol{\beta}_k, \sigma_k^2), \tag{19}$$

with

$$Q_1(\mathbf{w}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_{ik}(\mathbf{w}), \tag{20}$$

and,

$$
\begin{aligned}
Q_2(\boldsymbol{\beta}_k, \sigma_k^2) &= \sum_{i=1}^{n} \tau_{ik}^{(m)} \log \mathcal{N}\left(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2\right) \\
&= -\frac{1}{2} \left[\frac{1}{\sigma_k^2} \sum_{i=1}^{n} \tau_{ik}^{(m)} \left(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i\right)^2 + n_k^{(m)} \log \sigma_k^2 \right] \\
&\quad - \frac{n_k^{(m)}}{2} \log 2\pi \; ; \; k = 1, \ldots, K,
\end{aligned}
\tag{21}
$$

where $n_k^{(m)} = \sum_{i=1}^{n} \tau_{ik}^{(m)}$ can be interpreted as the number of points of the component k estimated at the iteration m. Thus, the maximization of Q can be performed by separately maximizing $Q_1(\mathbf{w})$ with respect to $\mathbf{w}$ and $Q_2(\boldsymbol{\beta}_k, \sigma_k^2)$ with respect to $(\boldsymbol{\beta}_k, \sigma_k^2)$ for all $k = 1, \ldots, K$. Maximizing Q_2 with respect to $\boldsymbol{\beta}_k$ consists in analytically solving a weighted least-squares problem. The estimates are given by:

$$
\begin{aligned}
\boldsymbol{\beta}_k^{(m+1)} &= \arg\min_{\boldsymbol{\beta}_k} \sum_{i=1}^{n} \tau_{ik}^{(m)} (x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2 \\
&= (\mathbf{T}^T \boldsymbol{W}_k^{(m)} \mathbf{T})^{-1} \mathbf{T}^T \boldsymbol{W}_k^{(m)} \mathbf{x},
\end{aligned}
\tag{22}
$$

where $\boldsymbol{W}_k^{(m)}$ is the $[n \times n]$ diagonal matrix of weights whose diagonal elements are $(\tau_{1k}^{(m)}, \ldots, \tau_{nk}^{(m)})$ and $\mathbf{T}$ is the $[n \times (p+1)]$ regression matrix.

Maximizing Q_2 with respect to σ_k^2 provides the following updating formula:

$$
\begin{aligned}
\sigma_k^{2(m+1)} &= \arg\min_{\sigma_k^2} \left[\frac{1}{\sigma_k^2} \sum_{i=1}^{n} \tau_{ik}^{(m)} \left(x_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{r}_i \right)^2 + n_k^{(m)} \log \sigma_k^2 \right] \\
&= \frac{1}{n_k^{(m)}} \sum_{i=1}^{n} \tau_{ik}^{(m)} (x_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{r}_i)^2.
\end{aligned}
\tag{23}
$$

The maximization of Q_1 with respect to $\mathbf{w}$ is a multinomial logistic regression problem weighted by $\tau_{ik}^{(m)}$ which we solve with a multi-class Iterative Reweighted Least Squares (IRLS) algorithm (Green, 1984; Chen et al., 1999; Krishnapuram et al., 2005; Chamroukhi et al., 2009).

It can be easily verified that the proposed algorithm is performed with a time complexity of $O(IJK^3p^2n)$, where I is the number of iterations of the EM algorithm and J is the average number of iterations required by its internal IRLS algorithm.

4.5. Denoising and segmenting a time series

In addition to performing time series parametrization, the proposed approach can be used to denoise and segment time series (or signals). The denoised time series can be approximated by the expectation $E(\mathbf{x}; \hat{\boldsymbol{\theta}}) = \big(E(x_1; \hat{\boldsymbol{\theta}}), \ldots, E(x_n; \hat{\boldsymbol{\theta}})\big)$ where

$$
\begin{aligned}
E(x_i; \hat{\boldsymbol{\theta}}) &= \int_{\mathbb{R}} x_i p(x_i; \hat{\boldsymbol{\theta}}) dx_i \\
&= \sum_{k=1}^{K} \pi_{ik}(\hat{\mathbf{w}}) \int_{\mathbb{R}} x_i \mathcal{N}\big(x_i; \hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i, \hat{\sigma}_k^2\big) dx_i \\
&= \sum_{k=1}^{K} \pi_{ik}(\hat{\mathbf{w}}) \hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i \ , \forall i = 1, \ldots, n,
\end{aligned}
\tag{24}
$$

and $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{w}}, \hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_K, \hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2)$ is the parameter vector obtained at convergence of the algorithm. The matrix formulation of the approximated signal $\hat{\mathbf{x}} = E(\mathbf{x}; \hat{\boldsymbol{\theta}})$ is given by:

$$
\hat{\mathbf{x}} = \sum_{k=1}^{K} \hat{\boldsymbol{\Pi}}_k \mathbf{T} \hat{\boldsymbol{\beta}}_k,
\tag{25}
$$

14

where $\hat{\boldsymbol{\Pi}}_k$ is a diagonal matrix whose diagonal elements are the proportions $(\pi_{1k}(\hat{\mathbf{w}}), \ldots, \pi_{nk}(\hat{\mathbf{w}}))$ associated with the kth regression model. On the other hand, a signal segmentation can also be obtained by computing the estimated label $\hat{z}_i$ of x_i according to the following rule:

$$\hat{z}_i = \arg \max_{1 \leq k \leq K} \pi_{ik}(\hat{\mathbf{w}}) \ , \quad \forall i = 1, \ldots, n. \tag{26}$$

Applying this rule guarantees the time series are segmented into contiguous segments if the probabilities π_{ik} are computed with a dimension $q = 1$ of $\boldsymbol{w}_k$; $k = 1, \ldots, K$.

4.6. Model selection

In a general application of the proposed model, the optimal values of (K, p, q) can be computed by using the Bayesian Information Criterion (BIC) (Schwarz, 1978) which is a penalized likelihood criterion, defined by

$$\text{BIC}(K, p, q) = L(\hat{\boldsymbol{\theta}}; \mathbf{x}) - \frac{\nu(K, p, q) \log(n)}{2} \ , \tag{27}$$

where $\nu(K, p, q) = K(p + q + 3) - (q + 1)$ is the number of parameters of the model and $L(\hat{\boldsymbol{\theta}}; \mathbf{x})$ is the log-likelihood obtained at convergence of the EM algorithm.

5. Experimental study using simulated signals

This section is devoted to an evaluation of the signal modeling performed by the proposed algorithm using simulated datasets. For this purpose, the proposed approach was compared with the piecewise regression and the Hidden Markov Regression approaches.

5.1. Evaluation criteria

Two evaluation criteria were used in the simulations. The first criterion is the mean square error between the true simulated curve without noise (which is the true denoised signal) and the estimated denoised signal given by:

- $\hat{x}_i = \sum_{k=1}^{K} \pi_{ik}(\hat{\mathbf{w}}) \hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i$ for the proposed model;

- $\hat{x}_i = \sum_{k=1}^{K} \hat{z}_{ik} \hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i$ for the piecewise polynomial regression model;

15

- $\hat{x}_i = \sum_{k=1}^{K} \omega_{ik}(\hat{\mathbf{\Psi}})\hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i$ for the HMM regression model.

This error criterion is computed by the formula $\frac{1}{n}\sum_{i=1}^{n}[E(x_i; \boldsymbol{\theta}) - \hat{x}_i]^2$. It is used to assess the models with regard to signal denoising and is called the denoising error.

The second criterion is the misclassification error rate between the simulated and the estimated partitions. It is used to assess the models with regard to signal segmentation. Note that other comparisons between the proposed approach and two versions of the piecewise polynomial regression approach including the running time can be found in (Chamroukhi et al., 2009).

5.2. Simulation protocol

The signals were simulated with the proposed regression model with hidden logistic process and all the simulations were performed for a number of segments $K = 3$. We chose the value $q = 1$ which guarantees a segmentation into contiguous intervals for the proposed model. We considered that all the time series were observed over 5 seconds with a constant sampling period ($\Delta t = t_i - t_{i-1}$ is constant).

Three experiments were performed:

- the first aims to observe the effect of the smoothness level of transitions on quality estimation. For this purpose two situations of simulated times series of $n = 300$ observations were considered. For the first situation, the time series consisted of three constant polynomial regimes ($K = 3, p = 0$) with a uniform noise level $\sigma = 1$. For the second situation, the time series consisted of three polynomial regimes of order 2 ($K = 3, p = 2$) with $n = 300$ and $\sigma = 0.5$. The set of simulation parameters for the two situations is given in Table 1. The smoothness level of transitions was tuned by means of the term $\lambda_k = \boldsymbol{w}_{k1}; k = 1, \ldots, K$, seen in section 4.1 and Fig. 3 (a). We used 10 smoothness levels for each situation. Fig. 4 shows the true denoised curves for situation 1 and situation 2, for the decreasing values of $|\lambda_k|$ shown in Table 2.

- the second aims to observe the effect of the sample size n on estimation quality. The sample size varied from 100 to 1000 is steps of 100, and the the values of the σ_k were set to $\sigma_1 = 1$, $\sigma_2 = 1.25$, and $\sigma_3 = 0.75$. Fig. 5 shows an example of simulated signal for $n = 700$.

16

- the third aims to observe the effect of the noise level σ. The noise level σ was assumed to be uniform for all the segments and varied from 0.5 to 5 is steps of 0.5, and the sample size was set to $n = 500$.

For each value of n, each value of σ and each value of the smoothness level of transitions we generated 20 samples and the values of assessment criteria were averaged over the 20 samples.

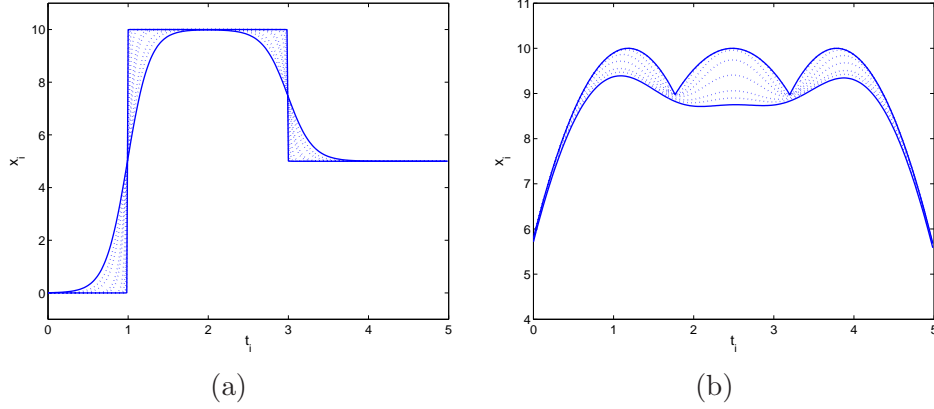| Situation 1 | $\boldsymbol{\beta}_1 = 0$ | $\boldsymbol{w}_1 = [3341.33, -1706.96]$ |
	$\boldsymbol{\beta}_2 = 5$	$\boldsymbol{w}_2 = [2436.97, -810.07]$
	$\boldsymbol{\beta}_3 = 10$	$\boldsymbol{w}_3 = [0, 0]$
Situation 2	$\boldsymbol{\beta}_1 = [-0.64, 14.4, -6]$	$\boldsymbol{w}_1 = [3767.58, -1510.19]$
	$\boldsymbol{\beta}_2 = [-21.25, 25, -5]$	$\boldsymbol{w}_2 = [2468.99, -742.55]$
	$\boldsymbol{\beta}_3 = [-78.64, 45.6, -6]$	$\boldsymbol{w}_3 = [0, 0]$

Table 1: Simulation parameters

Figure 4: The true denoised signals from abrupt transitions to smooth transitions for situation 1 (a) and situation 2 (b).

5.3. Initialization strategies and stopping rules

The proposed algorithm and the Hidden Markov regression algorithm were initialized as follows:

- In the proposed model **w** was set to the null vector;

- In the HMRM the initial probabilities were set to $\pi = (1, 0, \ldots, 0)$ and $A_{\ell k} = 0.5$ for $\ell \leq k \leq \ell + 1$;

17

Smoothness level of transitions	1	2	3	4	5	6	7	8	9	10
(a) $|\lambda_k|$ divided by:	1	2	5	10	20	40	50	80	100	125
(b) $|\lambda_k|$ divided by:	1	10	50	100	150	200	250	275	300	400

Table 2: The different smoothness levels from abrupt transitions to smooth transitons for the situations shown in Fig. 4.
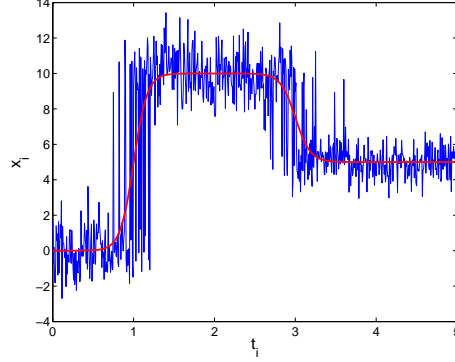

Figure 5: Example of simulated signal (with and without noise) for $n = 700$ and $\sigma = 1$ for situation 1 with a smoothness level of transtion corresponding to the level 8 in Table 2.

- to initialize $\boldsymbol{\beta}_k$ and σ_k^2, for $k = 1, \ldots, K$, several random segmentations of the signal into K segments were used as well as a uniform segmentation. On each segment k we fitted a polynomial regression model and then deduced the valued $\boldsymbol{\beta}_k$ and σ_k^2. The solution providing the highest likelihood was chosen.

The two algorithms were stopped when the relative variation of the log-likelihood function between two iterations $|\frac{L^{(m+1)} - L^{(m)}}{L^{(m)}}|$ was below 10^{-6} or after 1500 iterations.

5.4. Simulation results

Fig. 6 shows the denoising error and the misclassification error rate in relation to the smoothness level of transitions for the first situation (left) and for the second situation (right). It can be seen that the proposed approach performs the signals segmentation and denoising better than the piecewise regression and the HMRM approaches. While the results are closely similar when the transtions are abrupt (until level 3), the the proposed approach pro-

18

vides more accurate results than the two alternatives for smooth transitions for the two situations.
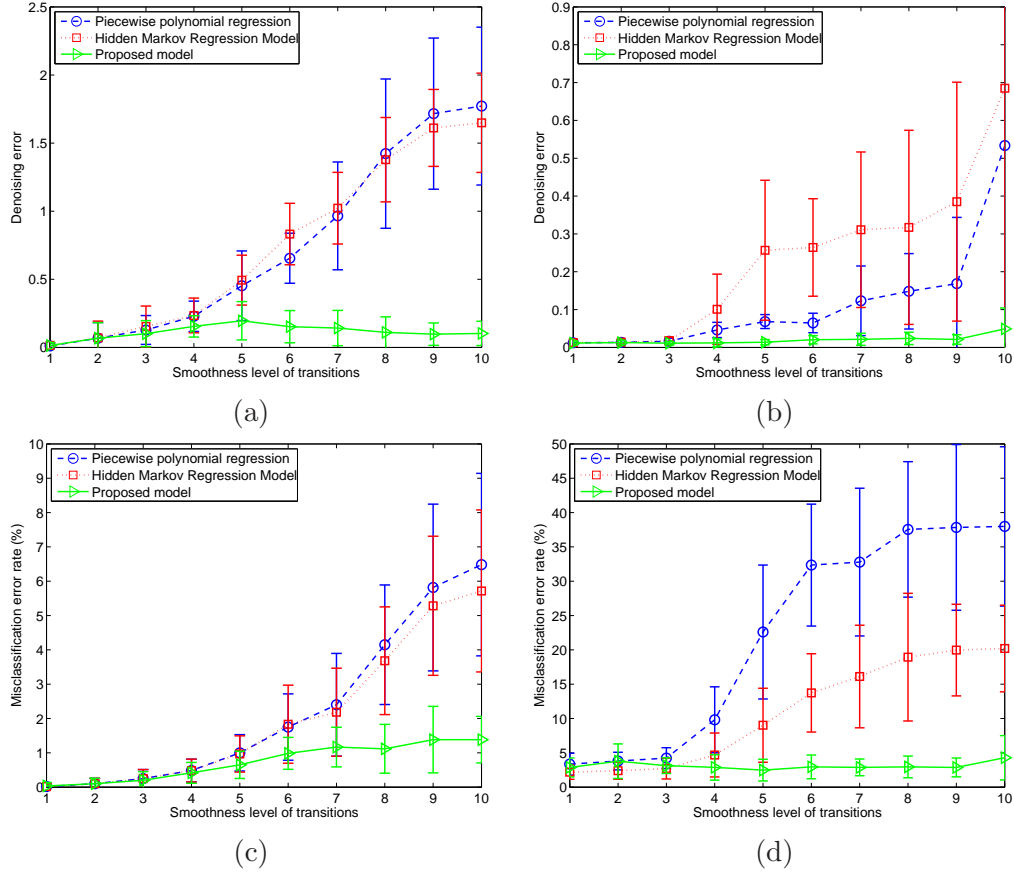

Figure 6: Denoising error (top) and misclassification error rate (bottom) with the error bars in the range of errors standard deviation, in relation to the smoothness level of transitions, obtained with the proposed approach (triangle), the piecewise polynomial regression approach (circle) and the HMRM approach (square) for the first situation (left) and for the second situation (right).

Fig. 7 shows the denoising error and the misclassification error rate in relation to the sample size n and the noise level σ. It can be seen in Fig. 7 (a) and Fig. 7 (b) that the segmentation error decreases when the sample size n increases for the proposed model which provides more accurate results than piecewise and the HMRM approaches. Fig. 7 (c) and Fig. 7 (d) show that when the noise level increases the proposed approach provides more stable results than to the two other alternative approaches.
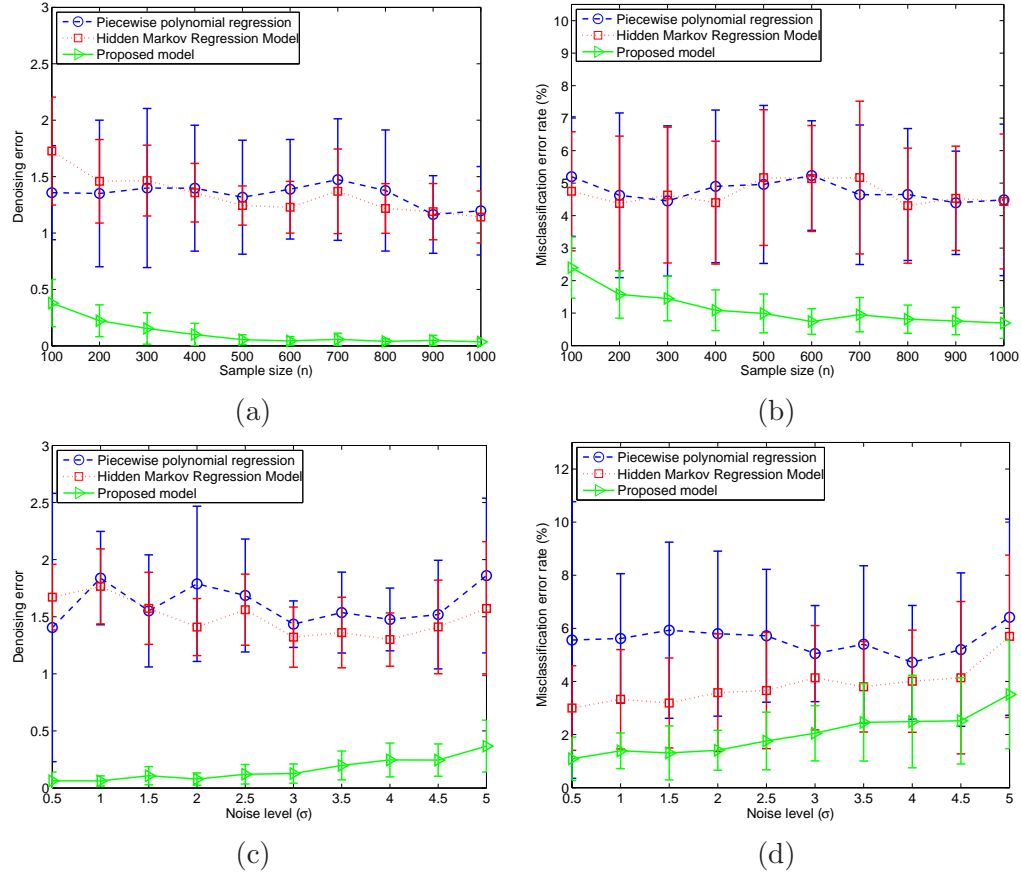
19

Figure 7: Denoising error (left) and misclassification error rate (right) with the error bars in the range of errors standard deviation, in relation to the sample size n for $(\sigma_1 = 1, \sigma_2 = 1.25, \sigma_3 = 0.75$ (a,b) and the noise level σ for $n = 500$ (c,d), obtained with the proposed approach (triangle), the piecewise polynomial regression approach (circle) and the HMRM approach (square).

20

6. Application to real signals

This section presents the results obtained by the proposed approach for the switch operation signals modeling and classification. Several types of signals were considered (with and without defects). The number of regression components was chosen in accordance with the number of electromechanical phases of a switch operation ($K = 5$). The value of q was set to 1, which guarantees segmentation into contiguous intervals for the proposed approach, and the degree of the polynomial regression p was set to 3 which is appropriate for the different regimes in the signals.

6.1. Real signal modeling

The proposed regression approach were applied to real signals of switch operations.

Fig. 8 (top) shows the original signals and the denoised signals (the denoised signal provided by the proposed approach is given by equation (25)). Fig. 8 (bottom) shows the variation of the probabilities π_{ik} over time. It can be seen that these probabilities are very closed to 1 when the kth regression model seems to be the most faithful to the original signal.

6.2. Real signal classification

This part is devoted to an evaluation of the classification accuracy of the proposed approach. A database of $N = 119$ real signals with known classes was used. This database was divided into two groups: a training base of 84 signals for learning the classes parameters and a test base of 35 signals for evaluating the classifier. The three parametrization methods were applied to applied to all the signals of the database, and the estimated parameters provided by each approach were used as the signal feature vector. After the parametrization step, the MDA was applied to the features extracted from the signals in the training data set. After the learning step, each signal, represented by its feature vector was classified using the Maximum A Posteriori (MAP) rule.

Three different classes of signals indexed by $g = 1, .., 3$, corresponding to the different operating states of the switch mechanism were considered. Thus, the considered classes were

- $g = 1$: no defect class;
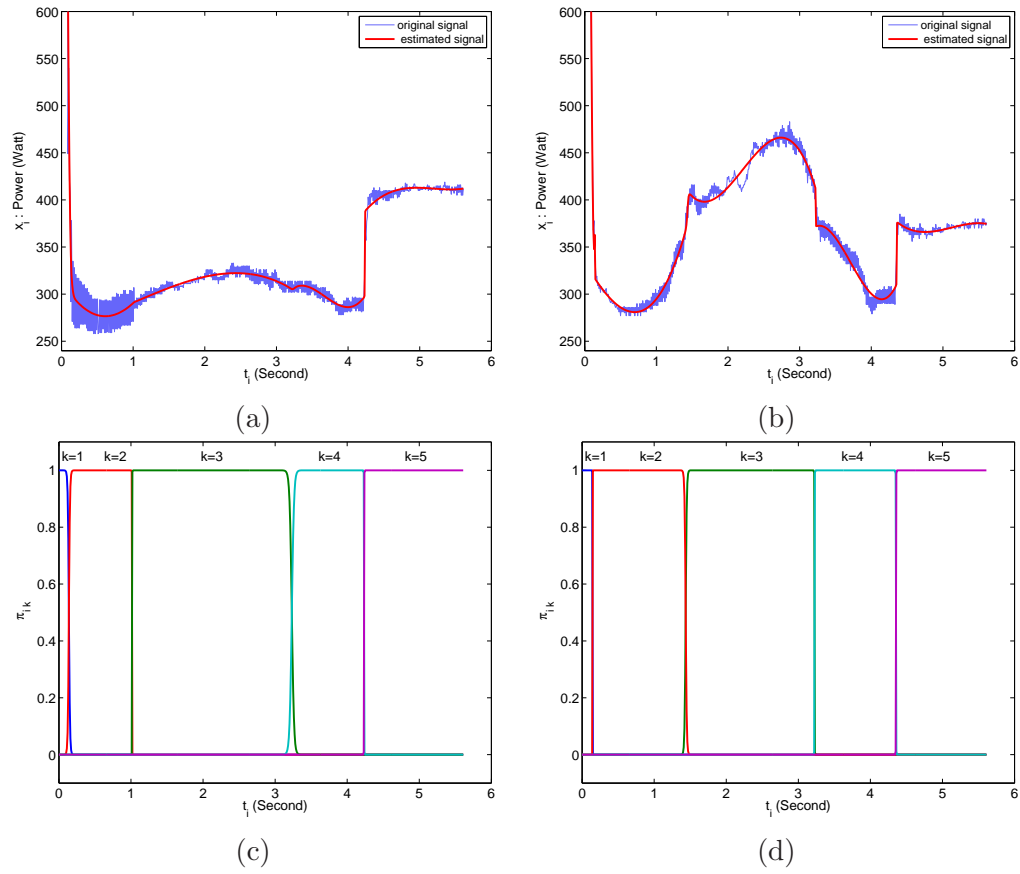
- $g = 2$: minor defect class;

21

Figure 8: Results obtained with the proposed approach for a signal without defect (a) and a signal with defect (b) with the original signal (in blue) and the estimated signal (in red) and the proportions π_{ik}, $k = 1, \ldots, 5$ for the estimated polynomial regression components over time (c) and (d).

22

- $g = 3$: critical defect class.

In what follows we shall use $\boldsymbol{y}_j$ to denote the feature vector $\boldsymbol{\theta}$ extracted from the signal $\mathbf{x}_j$, where the index $j = 1, \ldots, N$ corresponds to the signal number.

6.2.1. Modeling the operating classes with mixture models

Given a labelled collection of extracted features, the parameters of each class are learned using the Mixture Discriminant Analysis (MDA) Hastie and Tibshirani (1996). In this approach, the density of each class $g = 1, .., G$ with $G = 3$ is modeled by a Gaussian mixture distribution Hastie and Tibshirani (1996); McLachlan and Peel. (2000) defined by

$$p(\boldsymbol{y}_j | C_j = g; \boldsymbol{\Theta}_g) = \sum_{r=1}^{R_g} \alpha_{gr} \mathcal{N}\left(\boldsymbol{y}_j; \boldsymbol{\mu}_{gr}, \Sigma_{gr}\right), \tag{28}$$

where C_j is the discrete variable which takes its value in the set $\{1, \ldots, 3\}$ representing the class of the signal $\mathbf{x}_j$,

$$\boldsymbol{\Theta}_g = \left(\alpha_{g1}, \ldots, \alpha_{gR_g}, \boldsymbol{\mu}_{g1}, \ldots, \boldsymbol{\mu}_{gR_g}, \ldots, \Sigma_{g1}, \ldots, \Sigma_{gR_g}\right)$$

is the parameter vector of the mixture density of the class g with R_g is the number of mixture components and the α_{gr} $(r = 1, \ldots, R_g)$ are the mixing proportions satisfying $\sum_{r=1}^{R_g} \alpha_{gr} = 1$. The optimal number of Gaussian distributions R_g for each class g is computed by maximizing the BIC criterion Schwarz (1978):

$$\text{BIC}(R_g) = L(\hat{\boldsymbol{\Theta}}_g) - \frac{\nu_{R_g}}{2} \log(n_g), \tag{29}$$

where $\hat{\boldsymbol{\Theta}}_g$ is the maximum likelihood estimate of $\boldsymbol{\Theta}_g$ provided by the EM algorithm, ν_{R_g} is the dimension of the parameter vector $\boldsymbol{\Theta}_g$, and n_g is the cardinal number of class g.

Given the parameter vectors $\hat{\boldsymbol{\Theta}}_1$, $\hat{\boldsymbol{\Theta}}_2$, $\hat{\boldsymbol{\Theta}}_3$ estimated by the EM algorithm for the three classes of signals, each new signal designed by the feature vector $\boldsymbol{y}_j$ is assigned to the class $\hat{g}$ that maximizes the posterior probability that $\boldsymbol{x}_i$ belongs to the class g, with respect to $g = 1, \ldots, G$:

$$\hat{g} = \arg \max_{1 \leq g \leq G} p(C_j = g | \boldsymbol{y}_j; \hat{\boldsymbol{\Theta}}_g), \tag{30}$$

23

with

$$p(C_j = g | \boldsymbol{y}_j; \hat{\boldsymbol{\Theta}}_g) = \frac{p(C_j = g)p(\boldsymbol{y}_j | C_j = g; \hat{\boldsymbol{\Theta}}_g)}{\sum_{g'=1}^{G} p(C_j = g')p(\boldsymbol{x}_i | C_j = g'; \hat{\boldsymbol{\Theta}}_{g'})}, \qquad (31)$$

where $p(C_j = g)$ is the prior probability of the class g estimated by the proportion of the signals belonging to class g in the learning phase.

6.2.2. Classification results

The results in terms of correct classification rates are given in table (3) and the number of mixture components estimated by the BIC criterion for each class g, for the proposed modeling method, is given in table (4).

Modeling approach	Correct classification rate (%)
Piecewise regression model	83
HMRM	89
Proposed regression model	91

Table 3: Correct classification rates.

The correct classification rates show clearly that using the proposed regression approach for signals modeling show clearly that the proposed approach outperforms the two alternative approaches. The number of mixture

Class g	1	2	3
Number of mixture components R_g	1	1	3

Table 4: Number of mixture components selected with the BIC criterion.

components $R_g = 3$ selected with the BIC criterion for the third class (critical defect class) is attributed to the fact that this class contains signals covering a wide range of defects.

7. Conclusion

This paper proposes a new approach for time series modeling, in the context of the railway switch mechanism diagnosis. It is based on a regression model incorporating a discrete hidden logistic process. The logistic probability function used for the hidden variables allows smooth or abrupt transitions between various polynomial regression components over time. In addition to time series parametrization, the proposed model can provide accurate signals

24

segmentation and denoising. The performance of this approach in terms of signal modeling has been evaluated by comparing it to the piecewise polynomial regression approach and the Hidden Markov Regression Mode using simulated data and real data. Based on the proposed modeling approach, a mixture discriminant approach has been implemented to classify real signals.

Acknolegment

References

Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41:164–171.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Communications of the Association for Computing Machinery (CACM)*, 4(6):284.

Brailovsky, V. L. and Kempner, Y. (1992). Application of piecewise regression to detecting internal structure of signal. *Pattern recognition*, 25(11):1361–1370.

Chamroukhi, F., Samé, A., Govaert, G., and Aknin, P. (2009). A regression model with a hidden logistic process for feature extraction from time series. In *International Joint Conference on Neural Networks (IJCNN)*.

Chen, K., Xu, L., and Chi, H. (1999). Improved learning algorithms for mixture of experts in multiclass classification. *Neural Networks*, 12(9):1229–1252.

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

Ferrari-Trecate, G. and Muselli, M. (2002). A new learning method for piecewise linear regression. In *International Conference on Artificial Neural Networks (ICANN)*, pages 28–30.

Ferrari-Trecate, G., Muselli, M., Liberati, D., and Morari, M. (2002). A clustering technique for the identification of piecewise affine and hybrid systems. *Automatica*, 39:205–217.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, 53:789–798.

Fridman, M. (1993). Hidden markov model regression. Technical report, Institute of mathematics, University of Minnesota.

Green, P. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of The Royal Statistical Society, B*, 46(2):149–192.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society, B*, 58:155–176.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214.

Krishnapuram, B., Carin, L., Figueiredo, M., and Hartemink, A. (2005). Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968.

Lechevalier, Y. (1990). Optimal clustering on ordered set. *Technical report, The French National Institute for Research in Computer Science and Control (INRIA)*.

McGee, V. E. and Carleton, W. T. (1970). Piecewise regression. *Journal of the American Statistical Association*, 65:1109–1124.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

McLachlan, G. J. and Peel., D. (2000). *Finite mixture models*. New York: Wiley.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(730-738).

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Samé, A., Aknin, P., and Govaert, G. (2007). Classification automatique pour la segmentation des signaux unidimensionnels. *Rencontres de la SFC, ENST, Paris*.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Waterhouse, S. R. (1997). *Classification and regression using Mixtures of Experts*. PhD thesis, Department of Engineering, Cambridge University.