A regression model with a hidden logistic process for feature extraction from time serie

Faicel Chamroukhi, Allou Samé, Gérard Govaert, Patrice Aknin

A regression model with a hidden logistic process for feature extraction from time series

F. Chamroukhi[1,2], A. Samé[1], G. Govaert[2] and P. Aknin[1]

1- 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)

2- Compiègne University of Technology
HEUDIASYC Laboratory, UMR CNRS 6599
BP 20529, 60205 Compiègne Cedex (France)

***Abstract*— A new approach for feature extraction from time series is proposed in this paper. This approach consists of a specific regression model incorporating a discrete hidden logistic process. The model parameters are estimated by the maximum likelihood method performed by a dedicated Expectation Maximization (EM) algorithm. The parameters of the hidden logistic process, in the inner loop of the EM algorithm, are estimated using a multi-class Iterative Reweighted Least-Squares (IRLS) algorithm. A piecewise regression algorithm and its iterative variant have also been considered for comparisons. An experimental study using simulated and real data reveals good performances of the proposed approach.**

I. Introduction

IN the context of the predictive maintenance of the french railway switches (or points) which enable trains to be guided from one track to another at a railway junction, we have been brought to extract features from switch operations signals representing the electrical power consumed during a point operation (see Fig. 1). The final objective is to exploit these parameters for the identification of incipient faults.

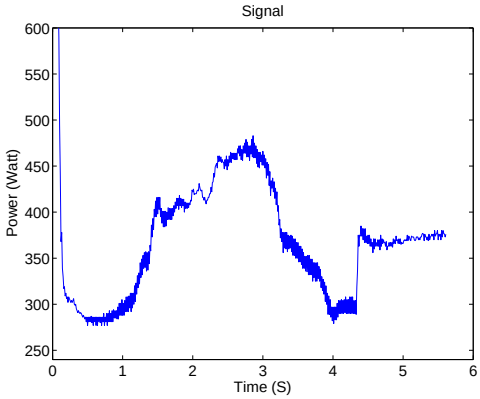


Fig. 1. Example of the electrical power consumed during a point operation.

The switch operations signals can be seen as time series presenting non-linearities and various changes in regime. Basic linear regression can not be adopted for this type of signals because a constant linear relationship is not adapted. As alternative to linear regression, some authors use approaches based on a piecewise regression model [14][17][18]. Piecewise regression is a segmentation method providing a partition of the data into K segments, each segment being characterized by its mean curve (constant, polynomial, ...) and its variance in the Gaussian case. Under this type of modeling, the parameters estimation is generally based on a global optimization using dynamic programming [2] like Fisher's algorithm [3]. This algorithm optimizes an additive criterion representing a cost function over all the segments of the signal [16][17]. However, the dynamic programming procedure is known to be computationally expensive. An iterative algorithm can be derived to improve the running time of Fisher's algorithm as in [19]. This iterative approach is a local optimization approach estimating simultaneously the regression model parameters and the transition points. These two approaches will be recalled in our work, where the second one will be extended to supposing different variances for the various segments instead of using a constant variance for all the segments. Other alternative approaches are based on Hidden Markov Models [9] in a context of regression [10] where the model parameters are estimated by the Baum-Welch algorithm [8].

The method we propose for feature extraction is based on a specific regression model incorporating a discrete hidden process allowing for abrupt or smooth transitions between various regression models. This approach has a connection with the switching regression model introduced by Quandt and Ramsey [13] and is very linked to the Mixture of Experts (ME) model introduced by Jordan and Jacobs [11] by the using of a time-dependent logistic transition function. The ME model, as discussed in [15], uses a conditional mixture modeling where the model parameters are estimated by the Expectation Maximization (EM) algorithm [1][5].

This paper is organized as follows. Section 2 recalls the piecewise regression model and two techniques of parameter estimation using a dynamic programming procedure: the method of global optimization of Fisher and its iterative variant. Section 3 introduces the proposed model and section 4 describes the parameters estimation via the EM algorithm. The fifth section is devoted to the experimental study using

simulated and real data.

II. Piecewise regression

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be n real observations of a signal or a time serie where x_i is observed at time t_i. The piecewise regression model supposes that the signal presents unknown transition points whose 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 $P_{n,K}$ of the time serie into K segments of lengths $n_1, \ldots, n_K$ such that:

$$P_{n,K} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}, \tag{1}$$

with $\boldsymbol{x}_k = \{x_i | i \in I_k\}$ and $I_k =]\gamma_k, \gamma_{k+1}]$.

Thus, the piecewise regression model generating the signal $\mathbf{x}$ is defined as follows:

$$\forall i = 1, \ldots, n, \quad x_i = \begin{cases} \boldsymbol{\beta}_1^T \boldsymbol{r}_i + \sigma_1 \epsilon_i & \text{if } i \in I_1 \\ \boldsymbol{\beta}_2^T \boldsymbol{r}_i + \sigma_2 \epsilon_i & \text{if } i \in I_2 \\ \vdots & \\ \boldsymbol{\beta}_K^T \boldsymbol{r}_i + \sigma_K \epsilon_i & \text{if } i \in I_K \end{cases}, \tag{2}$$

where $\boldsymbol{\beta}_k$, is the $(p+1)$-dimensional coefficients vector of a p degree polynomial associated to the k^{th} segment, $k \in \{1, \ldots, K\}$, $\boldsymbol{r}_i = (1, t_i, \ldots, (t_i)^p)^T$ is the time dependent $(p+1)$-dimensional covariate vector associated to the parameter $\boldsymbol{\beta}_k$ and the ϵ_i are independent random variables distributed according to a Gaussian distribution with zero mean and unit variance representing an additive noise on each segment k.

A. Parameter estimation

Under this model, the parameters estimation is performed by maximum likelihood. We assume a conditional independence of the data between the segments, and the data within a segment are also supposed to be conditionally independent. Thus, according to the model (2), the log-likelihood of the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1^2, \ldots, \sigma_K^2)$ and the transition points $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{K+1})$ characterizing the piecewise regression model is a sum of local log-likelihoods over all the segments and can be written as follows:

$$L(\boldsymbol{\psi}, \boldsymbol{\gamma}; \mathbf{x}) = \sum_{k=1}^{K} \ell_k(\boldsymbol{\beta}_k, \sigma_k^2; \boldsymbol{x}_k), \tag{3}$$

where

$$\begin{aligned} \ell_k(\boldsymbol{\beta}_k, \sigma_k^2; \boldsymbol{x}_k) &= \log p(\boldsymbol{x}_k; \boldsymbol{\beta}_k, \sigma_k^2) \\ &= \sum_{i \in I_k} \log \mathcal{N}(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2) \\ &= -\frac{1}{2} \sum_{i \in I_k} \Big[\log \sigma_k^2 + \frac{(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2}{\sigma_k^2} \Big] + \mathrm{c}_k, \end{aligned} \tag{4}$$

is the log-likelihood within the segment k and c_k is a constant. Thus, the log-likelihood is finally written as:

$$L(\boldsymbol{\psi}, \boldsymbol{\gamma}; \mathbf{x}) = -\frac{1}{2} \sum_{k=1}^{K} \sum_{i \in I_k} \Big[\log \sigma_k^2 + \frac{(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2}{\sigma_k^2} \Big] + \mathrm{C}, \tag{5}$$

where C is a constant.

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

$$\begin{aligned} J(\boldsymbol{\psi}, \boldsymbol{\gamma}) &= \sum_{k=1}^{K} \sum_{i \in I_k} \Big[\log \sigma_k^2 + \frac{(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2}{\sigma_k^2} \Big] \\ &= \sum_{k=1}^{K} J_k(\boldsymbol{\psi}, \gamma_k, \boldsymbol{\gamma}_{k+1}), \end{aligned} \tag{6}$$

where $J_k(\boldsymbol{\psi}, \gamma_k, \boldsymbol{\gamma}_{k+1}) = \sum_{i=\gamma_k+1}^{\gamma_{k+1}} \Big[\log \sigma_k^2 + \frac{(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2}{\sigma_k^2} \Big]$.

B. Fisher's algorithm for estimating the parameters of a piecewise regression model

The optimization algorithm of Fisher is an algorithm based on dynamic programming, providing the optimal partition of the data by minimizing an additive criterion [3][17][16]. This algorithm minimizes the criterion J or equivalently minimizes, with respect to $\boldsymbol{\gamma}$, the criterion

$$\begin{aligned} C_K(\boldsymbol{\gamma}) &= \min_{\boldsymbol{\psi}} J(\boldsymbol{\psi}, \boldsymbol{\gamma}) \\ &= \sum_{k=1}^{K} \min_{\boldsymbol{\beta}_k, \sigma_k^2} \sum_{i=\gamma_k+1}^{\gamma_{k+1}} \Big[\log \sigma_k^2 + \frac{(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2}{\sigma_k^2} \Big], \\ &= \sum_{k=1}^{K} c(\gamma_k, \gamma_{k+1}), \end{aligned} \tag{7}$$

with $c(\gamma_k, \gamma_{k+1}) = \sum_{i=\gamma_k+1}^{\gamma_{k+1}} \Big[\log \hat{\sigma}_k^2 + \frac{(x_i - \hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i)^2}{\hat{\sigma}_k^2} \Big]$, where

$$\begin{aligned} \hat{\boldsymbol{\beta}}_k^T &= \arg\min_{\boldsymbol{\beta}_k} \sum_{i=\gamma_k+1}^{\gamma_{k+1}} (x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2 \\ &= (\Phi_k^T \Phi_k)^{-1} \Phi_k^T \boldsymbol{x}_k, \end{aligned} \tag{8}$$

$\Phi_k = [\boldsymbol{r}_{\gamma_k+1}, \ldots, \boldsymbol{r}_{\gamma_{k+1}}]^T$ being the regression matrix associated to $\boldsymbol{x}_k$, and

$$\hat{\sigma}_k^2 = \frac{1}{n_k} \sum_{i=\gamma_k+1}^{\gamma_{k+1}} (x_i - \hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i)^2, \tag{9}$$

n_k being the number of points of the segment k.

It can be observed that the criterion $C_K(\boldsymbol{\gamma})$ is a sum of cost $c(\gamma_k, \gamma_{k+1})$ over the K segments. Therefore, due to the additivity of this criterion, its optimization can be performed using a dynamic programming procedure [16][2]. Dynamic programming considers that an optimal partition of the data into K segments is the union of an optimal partition into $K-1$ segments and a set of one segment. By introducing the cost

$$C_k(a, b) = \sum_{\ell=1}^{k} \min_{(\boldsymbol{\beta}, \sigma^2)} \sum_{i=a+1}^{b} \Big[\log \sigma_k^2 + \frac{(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2}{\sigma_k^2} \Big], \tag{10}$$

with $0 \leq a < b \leq n$ and $k = 1, \ldots, K$, the dynamic programming optimization algorithm runs as follows:

1) Step 1. (Initialization): This step consists of computing the cost matrix $C_1(a, b)$ corresponding to one segment $]a, b]$ for $0 \leq a < b \leq n$. This cost matrix is computed as follows:

$$\begin{aligned} C_1(a, b) &= \min_{(\boldsymbol{\beta}, \sigma^2)} \sum_{i=a+1}^{b} \Big[\log \sigma^2 + \frac{(x_i - \boldsymbol{\beta}^T \boldsymbol{r}_i)^2}{\sigma^2} \Big] \\ &= \sum_{i=a+1}^{b} \Big[\log \hat{\sigma}^2 + \frac{(x_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{r}_i)^2}{\hat{\sigma}^2} \Big], \end{aligned} \tag{11}$$

where $\hat{\boldsymbol{\beta}}^T$ and $\hat{\sigma}^2$ are computed respectively according to the equations (8) and (9) by replacing $]\gamma_k, \gamma_{k+1}]$ by $]a, b]$.

2) Step 2. (Dynamic programming procedure): This step consists of computing the optimal cost $C_k(a,b)$ for $k = 2,\ldots,K$ and $0 \leq a < b \leq n$ using the following formula:

$$C_k(a,b) = \min_{a \leq h \leq b} \Big[C_{k-1}(a,h) + C_1(h+1,b) \Big]. \qquad (12)$$

3) Step 3. (Finding the optimal partition): From the optimal costs $C_k(a,b)$, the optimal partition can be deduced (for more details see appendix A in [17]).

While the Fisher algorithm provides the global optimum, it is known to be computationally expensive. To accelerate the convergence of this algorithm, one can derive an iterative variant as in [19].

C. Iterative version of Fisher's algorithm

In the iterative procedure, the criterion $J(\psi,\gamma)$ given by equation (6) is iteratively minimized by starting from an initial value of the transition points $\gamma^{(0)} = (\gamma_1^{(0)},\ldots,\gamma_{K+1}^{(0)})$ and alternating the two following steps until convergence:

1) Regression step (at iteration m): Compute the regression model parameters $\psi^{(m)} = \{\boldsymbol{\beta}_k^{(m)}, \sigma_k^{2(m)}; k = 1\ldots,K\}$ for the current values of the transition points $\gamma^{(m)}$ by minimizing the criterion $J(\psi,\gamma^{(m)})$ given by equation (6) with respect to ψ. This minimization consists of performing K separated polynomial regressions and provides the following estimates:

$$\boldsymbol{\beta}_k^{T(m)} = (\Phi_k^{T(m)}\Phi_k^{(m)})^{-1}\Phi_k^{T(m)}\boldsymbol{x}_k^{(m)}, \qquad (13)$$

where $\Phi_k^{(m)} = [\boldsymbol{r}_{\gamma_k^{(m)}+1},\ldots,\boldsymbol{r}_{\gamma_{k+1}^{(m)}}]^T$ is the regression matrix associated to the elements of the k^{th} segment $\boldsymbol{x}_k^{(m)} = \{x_i | i \in]\gamma_k^{(m)}, \gamma_{k+1}^{(m)}]\}$ at the iteration m,

$$\sigma_k^{2(m)} = \frac{1}{n_k^{(m)}} \sum_{i=\gamma_k^{(m)}+1}^{\gamma_{k+1}^{(m)}} (x_i - \hat{\boldsymbol{\beta}}_k^{T(m)})^2. \qquad (14)$$

2) Segmentation step (at iteration m): Compute the transition points $\boldsymbol{\gamma}^{(m+1)} = (\gamma_1^{(m+1)},\ldots,\gamma_{K+1}^{(m+1)})$ by minimizing the criterion $J(\psi,\gamma)$ for the current value of $\psi = \psi^{(m)}$, with respect to $\boldsymbol{\gamma}$. This minimization can be performed using a dynamic programming procedure since the criterion $J(\psi^{(m)},\boldsymbol{\gamma})$ is additive. However, in contrast with the previous method, where the computation of the cost matrix $C_1(a,b)$ requires the computation of the regression model parameter $\{\hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k^2; k = 1,\ldots,K\}$ for $0 \leq a < b \leq n$, this iterative procedure simply uses the cost matrix computed with the current values of the parameters $\{\boldsymbol{\beta}_k^{T(m)}, \sigma_k^{2(m)}; k = 1\ldots,K\}$ which improves the running time of the algorithm.

The next section presents the proposed regression model with a hidden logistic process.

III. Regression model with a hidden logistic process

A. The global regression model

We represent a signal by the random sequence $\mathbf{x} = (x_1,...,x_n)$ of n real observations, where x_i is observed at time t_i. This sample is assumed to be generated by the following regression model with a discrete hidden logistic process $\mathbf{z} = (z_1,\ldots,z_n)$, where $z_i \in \{1,\ldots,K\}$:

$$\forall i = 1,\ldots,n, \quad \begin{cases} x_i = \boldsymbol{\beta}_{z_i}^T \boldsymbol{r}_i + \sigma_{z_i}\epsilon_i \\ \epsilon_i \sim \mathcal{N}(0,1) \end{cases} . \qquad (15)$$

In this model, $\boldsymbol{\beta}_{z_i}$ is the $(p+1)$-dimensional coefficients vector of a p degree polynomial, $\boldsymbol{r}_i = (1, t_i, \ldots, (t_i)^p)^T$ is the time dependent $(p+1)$-dimensional covariate vector associated to the parameter $\boldsymbol{\beta}_{z_i}$ and the ϵ_i are independent random variables distributed according to a Gaussian distribution with zero mean and unit variance. This model can be reformulated in a matrix form by

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{Z}_k(\mathbf{T}\boldsymbol{\beta}_k^T + \sigma_k\boldsymbol{\epsilon}), \qquad (16)$$

where $\mathbf{Z}_k$ is a diagonal matrix whose diagonal elements are $(z_{1k},\ldots,z_{nk})$ with $z_{ik} = 1$ if x_i is generated by the k^{th} regression model and 0 otherwise, $\mathbf{T} = \Big[\boldsymbol{r}_1,\ldots,\boldsymbol{r}_n\Big]^T$ is the $[n \times (p+1)]$ matrix of covariates, and $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^T$ is the noise vector distributed according to a zero mean multidimensional Gaussian density with identity covariance matrix.

B. 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 supposes 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

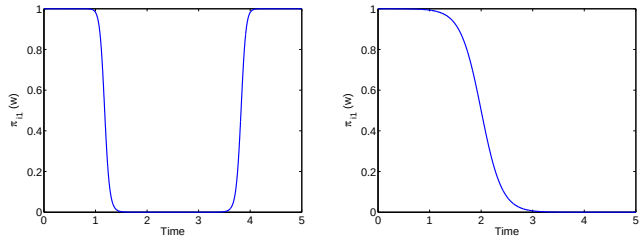
$$\pi_{ik}(\mathbf{w}) = p(z_i = k;\mathbf{w}) = \frac{\exp(\boldsymbol{w}_k^T\boldsymbol{v}_i)}{\sum_{\ell=1}^{K}\exp(\boldsymbol{w}_\ell^T\boldsymbol{v}_i)} \quad , \qquad (17)$$

is the logistic transformation of a linear function of the time-dependent covariate $\boldsymbol{v}_i = (1, t_i, \ldots, (t_i)^q)^T$, $\boldsymbol{w}_k = (\boldsymbol{w}_{k0},\ldots,\boldsymbol{w}_{kq})^T$ is the $(q+1)$-dimensional coefficients vector associated to 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:

$$p(\mathbf{z};\mathbf{w}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\frac{\exp(\boldsymbol{w}_k^T\boldsymbol{v}_i)}{\sum_{\ell=1}^{K}\exp(\boldsymbol{w}_\ell^T\boldsymbol{v}_i)}\right)^{z_{ik}} \quad , \qquad (18)$$

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

The pertinence of the logistic transformation in terms of flexibility of transition can be illustrated through simple examples with $K = 2$ components. 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 $\boldsymbol{w}_k$. In that 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 $\pi_{i2}(\mathbf{w}) = 1 - \pi_{i1}(\mathbf{w})$. As shown in Fig. 2, the dimension q controls the number of changes in the temporal variations 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 convex segments, the dimension q of $\boldsymbol{w}_k$ must be set to 1.

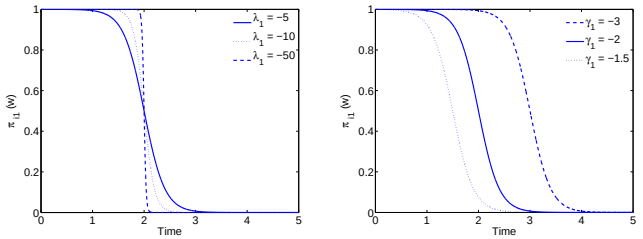


(a) (b)

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

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 $\boldsymbol{w}_k = \lambda_k(\gamma_k, 1)^T$ of $\boldsymbol{w}_k$, where $\lambda_k = \boldsymbol{w}_{k1}$ and $\gamma_k = \frac{\boldsymbol{w}_{k0}}{\boldsymbol{w}_{k1}}$. As it can be shown in Fig. 3 (a), the parameter λ_k controls the quality of transitions between classes, more the absolute value of λ_k is large, more the transition between the z_i is abrupt, while the parameter γ_k controls the transition time point by the means of the inflexion point of the curve (see Fig. 3 (b)). In that case of 2 classes and $q = 1$, the transition time point is the solution of $\boldsymbol{w}_{k0} + \boldsymbol{w}_{k1}t = 0$ which is $t = -\gamma_k$.



(a) (b)

Fig. 3. Variation of $\pi_{i1}(\mathbf{w})$ over time for a dimension $q = 1$ of $\boldsymbol{w}_k$ and (a) different values of $\lambda_k = \boldsymbol{w}_{k1}$ and (b) different values of $\gamma_k = \frac{\boldsymbol{w}_{k0}}{\boldsymbol{w}_{k1}}$.

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

C. The generative model of signals

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

- generate the hidden process $(z_1, \ldots, z_n)$ with $z_i \sim \mathcal{M}(1, \pi_{i1}(\mathbf{w}), \ldots, \pi_{iK}(\mathbf{w}))$,
- generate each observation x_i according to the Gaussian distribution $\mathcal{N}(\cdot; \boldsymbol{\beta}_{z_i}^T \boldsymbol{r}_i, \sigma_{z_i}^2)$.

IV. Parameter estimation

From the model (15), it can be proved that the random variable x_i is distributed according to the normal mixture density

$$p(x_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{w}) \mathcal{N}\left(x_i; \beta_k^T \boldsymbol{r}_i, \sigma_k^2\right) \ , \tag{19}$$

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

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{x}) &= \log \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_{ik}(\mathbf{w}) \mathcal{N}\left(x_i; \beta_k^T \boldsymbol{r}_i, \sigma_k^2\right). \end{aligned} \tag{20}$$

Since the direct maximization of this likelihood is not straightforward, we use the Expectation Maximization (EM) algorithm [1][5] to perform the maximization.

A. The dedicated EM algorithm

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

*1) **E Step (Expectation)**:* This step consists of 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):

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) &= E\left[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) | \mathbf{x}; \boldsymbol{\theta}^{(m)}\right] \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \left[\pi_{ik}(\mathbf{w}) \mathcal{N}\left(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2\right)\right], \end{aligned} \tag{21}$$

where

$$\begin{aligned} \tau_{ik}^{(m)} &= p(z_{ik} = 1 | x_i; \boldsymbol{\theta}^{(m)}) \\ &= \frac{\pi_{ik}(\mathbf{w}^{(m)}) \mathcal{N}(x_i; \boldsymbol{\beta}_k^{T(m)} \boldsymbol{r}_i, \sigma_k^{2(m)})}{\sum_{\ell=1}^{K} \pi_{i\ell}(\mathbf{w}^{(m)}) \mathcal{N}(x_i; \boldsymbol{\beta}_\ell^{T(m)} \boldsymbol{r}_i, \sigma_\ell^{2(m)})}, \end{aligned} \tag{22}$$

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

*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 $\boldsymbol{\theta}$. The maximization of Q can be performed by separately maximizing

$$Q_1(\mathbf{w}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(m)} \log \pi_{ik}(\mathbf{w}) \tag{23}$$

and, for all $k = 1, \ldots, K$

$$Q_2(\boldsymbol{\beta}_k, \sigma_k^2) = \sum_{i=1}^{n} \tau_{ik}^{(m)} \log \mathcal{N}(x_i; \boldsymbol{\beta}_k^T \boldsymbol{r}_i, \sigma_k^2) \tag{24}$$

Maximizing Q_2 with respect to the $\boldsymbol{\beta}_k$ consists of analytically solving a weighted least-squares problem. The estimates are straightforward and are as follows:

$$\begin{aligned} \boldsymbol{\beta}_k^{T(m+1)} &= \arg\min_{\boldsymbol{\beta}_k} \sum_{i=1}^{n} \tau_{ik}^{(m)} (x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2 \\ &= (\mathbf{T}^T \boldsymbol{W}_k^{(m)} \mathbf{T})^{-1} \mathbf{T}^T \boldsymbol{W}_k^{(m)} \mathbf{x}, \end{aligned} \tag{25}$$

with $\boldsymbol{W}_k^{(m)}$ is the $[n \times n]$ diagonal matrix of weights whose diagonal elements are $(\tau_{1k}^{(m)}, \ldots, \tau_{nk}^{(m)})$ and $\mathbf{x} = (x_1, \ldots, x_n)^T$ is the $[(n+1) \times 1]$ vector of observations.
Maximizing Q_2 with respect to the σ_k^2 provides the following estimates:

$$\begin{aligned} \sigma_k^{2(m+1)} &= \arg\min_{\sigma_k^2} \sum_{i=1}^{n} \tau_{ik}^{(m)} \left[\log \sigma_k^2 + \frac{(x_i - \boldsymbol{\beta}_k^T \boldsymbol{r}_i)^2}{\sigma_k^2}\right] \\ &= \frac{1}{\sum_{i=1}^{n} \tau_{ik}^{(m)}} \sum_{i=1}^{n} \tau_{ik}^{(m)} (x_i - \boldsymbol{\beta}_k^{T(m+1)} \boldsymbol{r}_i)^2. \end{aligned} \tag{26}$$

The maximization of Q_1 with respect to $\mathbf{w}$ is a multinomial logistic regression problem weighted by the $\tau_{ik}^{(m)}$. We use a multi-class Iterative Reweighted Least Squares (IRLS) algorithm [12][4][7] to solve it. The IRLS algorithm is detailed in the following section.

3) The Iteratively Reweighted Least Squares (IRLS) algorithm: The IRLS algorithm is used to maximize $Q_1(\mathbf{w})$ with respect to the parameter $\mathbf{w}$, in the M step, at each iteration m of the EM algorithm. To estimate the parameters vector $\mathbf{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$, since $\sum_{k=1}^{K} \pi_{ik}(\mathbf{w}) = 1$, $\boldsymbol{w}_K$ is set to the null vector to avoid the identification problems. The IRLS algorithm is equivalent to the Newton-Raphson algorithm, which consists of starting from a vector $\mathbf{w}^{(0)}$, and updating the estimation of $\mathbf{w}$ as follows:

$$\mathbf{w}^{(c+1)} = \mathbf{w}^{(c)} - \left[H(\mathbf{w}^{(c)})\right]^{-1} g(\mathbf{w}^{(c)}) \ , \tag{27}$$

where $H(\mathbf{w}^{(c)})$ and $g(\mathbf{w}^{(c)})$ are respectively the Hessian and the gradient of $Q_1(\mathbf{w})$ evaluated at $\mathbf{w} = \mathbf{w}^{(c)}$. In [4], authors use an approximation of the Hessian matrix to accelerate the convergence of the algorithm, while, in our case we use the exact Hessian matrix to perform well the maximum likelihood estimation as noticed in [7]. Since there are $K-1$ parameters vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{K-1}$ to be estimated, the Hessian matrix $H(\mathbf{w}^{(c)})$ consists of $(K-1)\times(K-1)$ block matrices $H_{k\ell}(\mathbf{w}^{(c)})(k, \ell = 1, \ldots, K-1)$ [7] where :

$$\begin{aligned} H_{k\ell}(\mathbf{w}^{(c)}) &= \frac{\partial^2 Q_1(\mathbf{w})}{\partial \boldsymbol{w}_k \partial \boldsymbol{w}_\ell}\Big|_{\mathbf{w}=\mathbf{w}^{(c)}} \\ &= -\sum_{i=1}^{n} \pi_{ik}(\mathbf{w}^{(c)})[\delta_{k\ell} - \pi_{i\ell}(\mathbf{w}^{(c)})]\boldsymbol{v}_i {\boldsymbol{v}_i}^T, \end{aligned} \quad (28)$$

where $\delta_{k\ell}$ is the kronecker symbol ($\delta_{k\ell}$ = 1 if $k = \ell$, 0 otherwise). The gradient of $Q_1(\mathbf{w})$ consists of $K-1$ gradients corresponding to the vectors $\boldsymbol{w}_k$ for $k = 1, \ldots, K-1$ and is computed as follows:

$$\begin{aligned} g(\mathbf{w}^{(c)}) &= \frac{\partial Q_1(\mathbf{w})}{\partial \mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}^{(c)}} \\ &= [g_1(\mathbf{w}^{(c)}), \ldots, g_{K-1}(\mathbf{w}^{(c)})]^T, \end{aligned} \quad (29)$$

with

$$\begin{aligned} g_k(\mathbf{w}^{(c)}) &= \frac{\partial Q_1(\mathbf{w})}{\partial \boldsymbol{w}_k}\Big|_{\mathbf{w}=\mathbf{w}^{(c)}} \\ &= \sum_{i=1}^{n} [\tau_{ik}^{(m)} - \pi_{ik}(\mathbf{w}^{(c)})]\boldsymbol{v}_i^T \; ; \; k = 1, \ldots, K-1. \end{aligned} \quad (30)$$

Applying algorithm (27) provides the parameter $\mathbf{w}^{(m+1)}$.

Algorithm (1) summarizes the proposed algorithm.

Algorithm 1 Pseudo code for the proposed algorithm.

Initialize:
fix a threshold $\epsilon > 0$; $m \leftarrow 0$ (iteration)
choose an initial $\boldsymbol{\theta}^{(m)}$=$\{\boldsymbol{w}_k^{(m)}, \boldsymbol{\beta}_k^{(m)}, \sigma_k^{2(m)}; k$=$1, \ldots, K\}$
Compute the initial value of $\pi_{ik}^{(m)}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ using equation (17)
while increment in log-likelihood $> \epsilon$ **do**
  {**E** step}: Compute the $\tau_{ik}^{(m)}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ using equation (22)
  {**M** step}: for $k = 1, \ldots, K$
  Compute $\boldsymbol{\beta}_k^{(m+1)}$ using equation (25)
  Compute $\sigma_k^{2(m+1)}$ using equation (26)
  compute $\mathbf{w}^{(m+1)}$ using the IRLS algorithm:
  {**IRLS** loop}:
  Initialize:
  set a threshold $\delta > 0$; $c \leftarrow 0$ (iteration)
  set $\mathbf{w}^{(c)} = \mathbf{w}^{(m)}$
  while increment in $Q_1(\mathbf{w}) > \delta$ **do**
    Compute $\pi_{ik}^{(c)}$ using equation (17)
    Compute $\mathbf{w}^{(c+1)}$ using equation (27)
    $c \leftarrow c + 1$
  end while
  $\mathbf{w}^{(m+1)} \leftarrow \mathbf{w}^{(c)}$
  $\pi_{ik}^{(m+1)} \leftarrow \pi_{ik}^{(c)}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$
  $m \leftarrow m + 1$
end while
$\hat{\boldsymbol{\theta}} = (\boldsymbol{w}_1^{(m)}, \ldots, \boldsymbol{w}_K^{(m)}, \boldsymbol{\beta}_1^{(m)}, \ldots, \boldsymbol{\beta}_K^{(m)}, \sigma_1^{2(m)}, \ldots, \sigma_K^{2(m)})$

B. Denoising and segmenting a signal

In addition to providing a signal parametrization, the proposed approach can be used to denoise and segment signals. The denoised signal 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

$$\begin{aligned} E(x_i; \hat{\boldsymbol{\theta}}) &= \int_{I\!R} x_i p(x_i; \hat{\boldsymbol{\theta}}) dx_i \\ &= \sum_{k=1}^{K} \pi_{ik}(\hat{\mathbf{w}}) \hat{\boldsymbol{\beta}}_k^T \boldsymbol{r}_i \;\; , \forall i = 1, \ldots, n, \end{aligned} \quad (31)$$

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 parameters vector obtained at the convergence of the algorithm. The matrix formulation of the approximated signal $\hat{\mathbf{x}} = E(\mathbf{x}; \boldsymbol{\theta})$ is given by:

$$\hat{\mathbf{x}} = \sum_{k=1}^{K} \hat{\mathbf{\Pi}}_k \mathbf{T} \hat{\boldsymbol{\beta}}_k^T, \quad (32)$$

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

$$\hat{z}_i = \arg \max_{1 \le k \le K} \pi_{ik}(\hat{\mathbf{w}}) \;\; , \quad \forall i = 1, \ldots, n. \quad (33)$$

C. Model selection

In a general use of the proposed model, the optimal values of (K, p, q) can be computed by using the Bayesian Information Criterion [6] which is a penalized likelihood criterion, defined by

$$BIC(K, p, q) = L(\hat{\boldsymbol{\theta}}; \mathbf{x}) - \frac{\nu(K, p, q) \log(n)}{2} \;\; , \quad (34)$$

where $\nu(K, p, p) = K(p + q + 3) - (q + 1)$ is the number of parameters of the model and $L(\hat{\boldsymbol{\theta}}; \mathbf{x})$ is the log-likelihood obtained at the convergence of the EM algorithm. If the goal is to segment the data into convex segments q must be set to 1.

V. Experiments

This section is devoted to the evaluation of the proposed algorithm using simulated and real data sets. For this purpose, the proposed approach is compared with the piecewise regression algorithm of Fisher and its iterative version. All the signals have been simulated from the piecewise regression model given by equation (2). Three evaluation criteria are used in the simulations.

- the first one is the misclassification rate between the simulated partition P and the estimated partition $\hat{P}$,
- the second one is the mean square error between the expectations computed with the true parameter $\boldsymbol{\theta}$ and the estimated parameter $\hat{\boldsymbol{\theta}}$: $\frac{1}{n}\sum_{i=1}^{n}[E(x_i; \boldsymbol{\theta}) - E(x_i; \hat{\boldsymbol{\theta}})]^2$ where $E(x_i; \hat{\boldsymbol{\theta}})$ is computed according to equation (32) for the proposed model, and $E(x_i; \hat{\boldsymbol{\theta}}) = \boldsymbol{\beta}_{\hat{z}_i}^T \boldsymbol{r}_i$ for the piecewise regression models. This error is used to asses the signal in terms of signal denoising and we call it the error of denoising.
- the third criterion is the running time.

A. Simulated signals

1) Protocol of simulations: For all the simulations, we set the number of segments (respectively the number of states of the hidden variable z_i for the proposed model) to $K = 3$ and the order of polynomial to $p = 2$. We choose the value $q = 1$ which guarantees a segmentation into contiguous intervals. We consider that all signals are observed over 5 seconds (the time interval being fixed to $[0, 5]$ Seconds) with a constant period of sampling $\Delta t = t_i - t_{i-1}$ depending on the sample size $n = 100, 200, ..., 1000$. For each size n we generate 20 samples. The values of assessment

criteria are averaged over the 20 samples. Two situations have been considered for simulations.

- situation1: the transition time points are set to $(0, 0.6, 4, 5)$ seconds, which correspond $\gamma_1 = 0$, $\gamma_2 = \frac{0.6}{\Delta t}$, $\gamma_3 = \frac{4}{\Delta t}$ and $\gamma_4 = \frac{5}{\Delta t}$. The set of parameters of simulations $\{\beta_k, \sigma_k^2; k = 1, \ldots, K\}$ corresponding to this situation is given by table I,
- situation2: the transition time points are set to $(0, 1, 3.5, 5)$ seconds, which correspond to $\gamma_1 = 0$, $\gamma_2 = \frac{1}{\Delta t}$, $\gamma_3 = \frac{3.5}{\Delta t}$ and $\gamma_4 = \frac{5}{\Delta t}$. The set of parameters of simulations $\{\beta_k, \sigma_k^2; k = 1, \ldots, K\}$ corresponding to this situation is given by table II.

Fig. 4 shows an example of simulated signals for the two situations.

$\boldsymbol{\beta}_1 = (735, -1320, 1000)^T$	$\sigma_1^2 = 4$
$\boldsymbol{\beta}_2 = (270, 60, -15)^T$	$\sigma_2^2 = 10$
$\boldsymbol{\beta}_3 = (320, 40, -4)^T$	$\sigma_3^2 = 15$

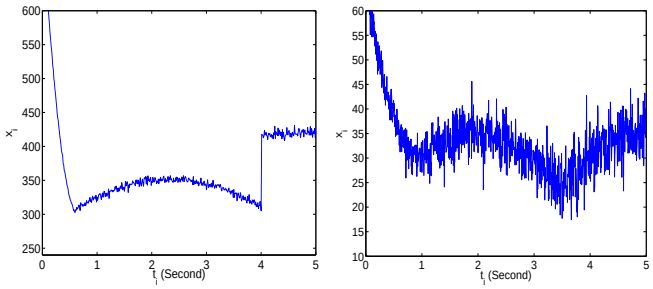
TABLE I

PARAMETERS OF SIMULATIONS FOR SITUATION 1.

$\boldsymbol{\beta}_1 = (65, -70, 35)^T$	$\sigma_1^2 = 4$
$\boldsymbol{\beta}_2 = (15, 20, -5)^T$	$\sigma_2^2 = 10$
$\boldsymbol{\beta}_3 = (-90, 50, -5)^T$	$\sigma_3^2 = 15$

TABLE II

PARAMETERS OF SIMULATIONS FOR SITUATION 2.



(a) (b)

Fig. 4. Simulated signal for the first situation (a) and the second situation (b) for $n = 1000$.

2) Strategy of initialization: The proposed algorithm is initialized as follows:

- $\boldsymbol{w}_k = (0, \ldots, 0)^T \ \forall k = 1, \ldots, K$,
- to initialize $\boldsymbol{\beta}_k$, we segment the signal uniformly into K segments and on each segment k we fit a regression model, characterized by $\boldsymbol{\beta}_k$,
- $\sigma_k^2 = 1$ for $k = 1, \ldots, K$.

For the iterative algorithm based on dynamic programming, several random initializations are used in addition to one initialization consisting of segmenting the signal into K uniform segments, and the best solution corresponding to the smallest value of the criterion $J(\boldsymbol{\psi}, \boldsymbol{\gamma})$ is then selected. In the random initializations, the condition that the transition points are ordered in the time is respected. The algorithm is stopped when the increment in the criterion $J(\boldsymbol{\psi}, \boldsymbol{\gamma})$ is below 10^{-6}.

B. Results

Fig. 5 (top) and Fig. 6 (top) show the misclassification rate in relation to the sample size n for the two situations of simulated data. It can be observed that the performance of the proposed approach in terms of classification is similar than the global optimization approach. Fig. 5 (down) and Fig. 6 (down) show the error of denoising. The low denoising error obtained by the proposed approach involves a good performance in terms of estimating the true model of the signal, compared to the piecewise regression approaches. Finally, Fig. 7 shows the slight variation of the running time of the proposed approach in relation to the sample size. The proposed algorithm is very fast compared to the two other approaches.

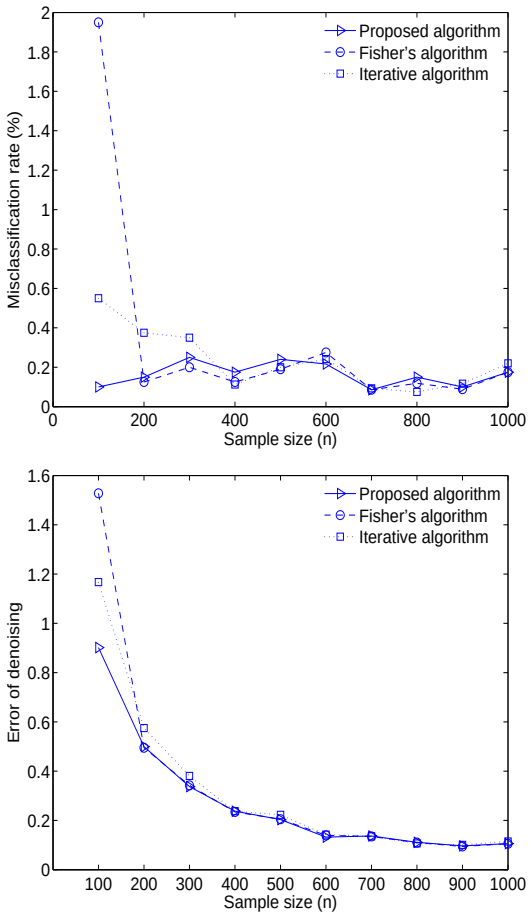


Fig. 5. Average misclassification rates (top) and average error of denoising (down) in relation to the sample size n obtained with the proposed approach (triangle), Fisher's algorithm (circle) and the iterative version of Fisher's algorithm (square) for the first situation.

C. Real signals

This section presents the results obtained by the proposed approach for signals of switch points operations. Two situations of signals have been considered: one without defect and one with a critical defect. The number of regressive components is chosen in accordance with the number of electromechanical phases of a switch points operation ($K = 5$). The value of q has been set to 1, which guarantees a segmentation into homogeneous intervals, and the degree of the polynomial regression p has been set to 3 which is adapted to the different regimes in the signals.

Fig. 8 (top) shows the original signals and the denoised signals (the denoised signal is given by equation (32)). Fig. 8 (middle) shows the variation of the probabilities π_{ik} over time. It can be observed that these probabilities are very closed to 1 when the k^{th} regressive model seems to be the most faithful to the original signal. The five regressive components involved in each signal are shown in Fig. 8 (down). Fig. 9 shows the segmentation, the estimated signals

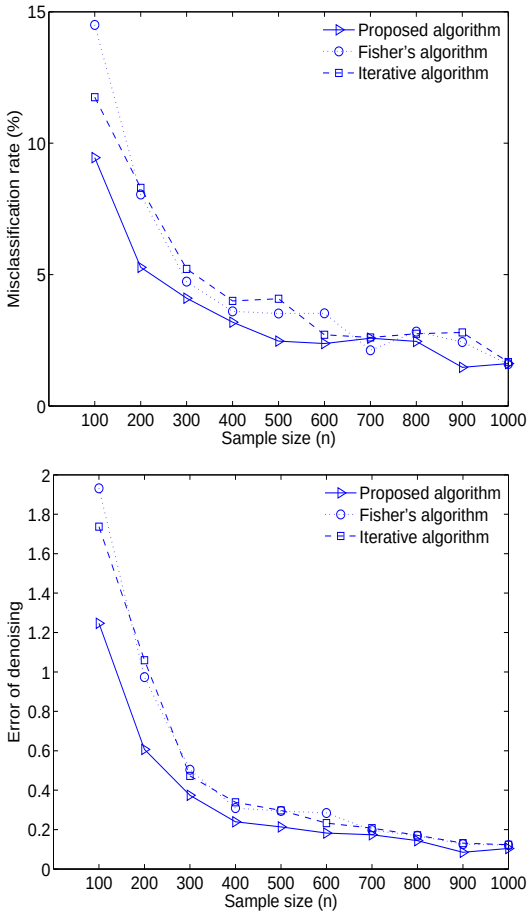


Fig. 6. Average misclassification rates (top) and average error of denoising (down) in relation to the sample size n obtained with the proposed approach (triangle), Fisher's algorithm (circle) and the iterative version of Fisher's algorithm (square) for the second situation.

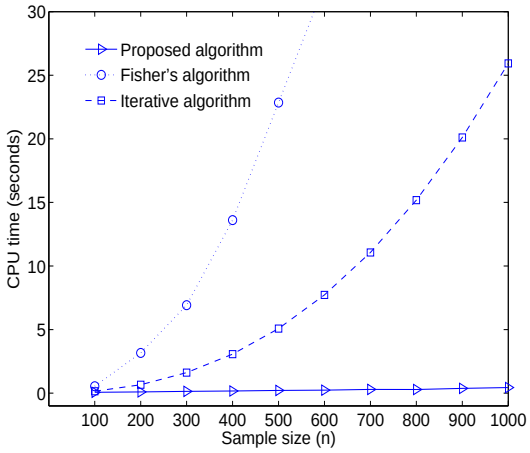


Fig. 7. Average running time in relation to the sample size n obtained with the proposed approach (triangle), Fisher's algorithm (circle) and the iterative version of Fisher's algorithm (square).

and and the Mean Square Errors (MSE) between the original signal and the estimated signal, obtained with the three methods for the two situations of signals.

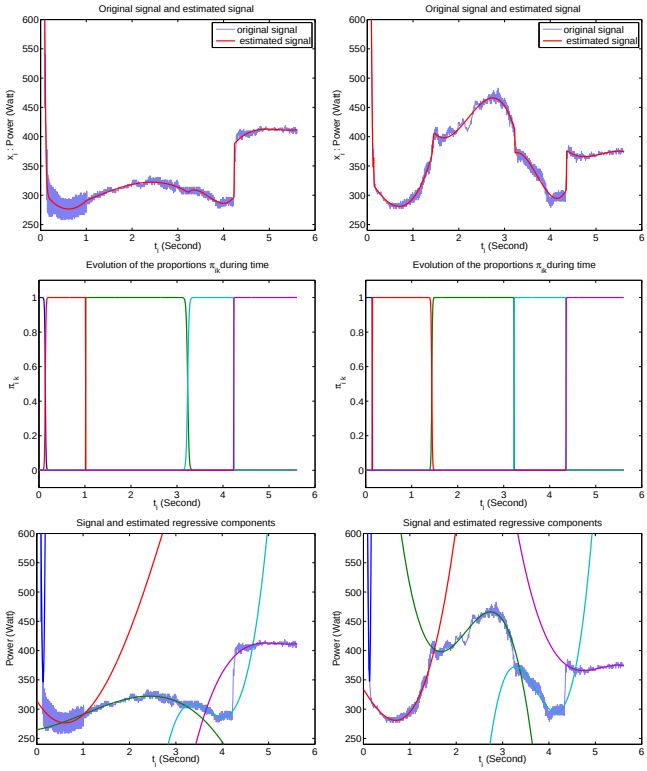


Fig. 8. Obtained results for a signal without defect (left) and for a signal with defect (right).

To illustrate the signal generation model, we generate two signals according to the proposed model using the parameters estimated by the EM algorithm. It can be seen that the generated signals are very similar to the original signals (see Fig. 10).

VI. Conclusion

In this paper a new approach for feature extraction from time series signals, in the context of the railway switch mechanism monitoring, has been proposed. This approach is based on a regression model incorporating a discrete hidden logistic process. The logistic probability function, used for the hidden variables, allows for smooth or abrupt transitions between polynomial regressive components over time. In addition to signals parametrization, an accurate denoising and segmentation of signals can be derived from the proposed model. The experiments applied to real and simulated data have shown good performances of the proposed approach compared to two algorithms devoted to the piecewise regression.

References

[1] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, B, 39(1): 1-38, 1977.

[2] R. Bellman, "On the approximation of curves by line segments using dynamic programming," *Communications of the Association for Computing Machinery (CACM)* (4), No. 6, pp. 284 June 1961.

[3] W. D. Fisher, "On grouping for maximum homogeneity," *Journal of American Statistics. Society* 53, 789-798, 1958.

[4] B. Krishnapuram, L. Carin, M.A.T. Figueiredo and A.J. Hartemink, "Sparse multinomial logistic regression: fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 27(6): 957-968, June 2005.

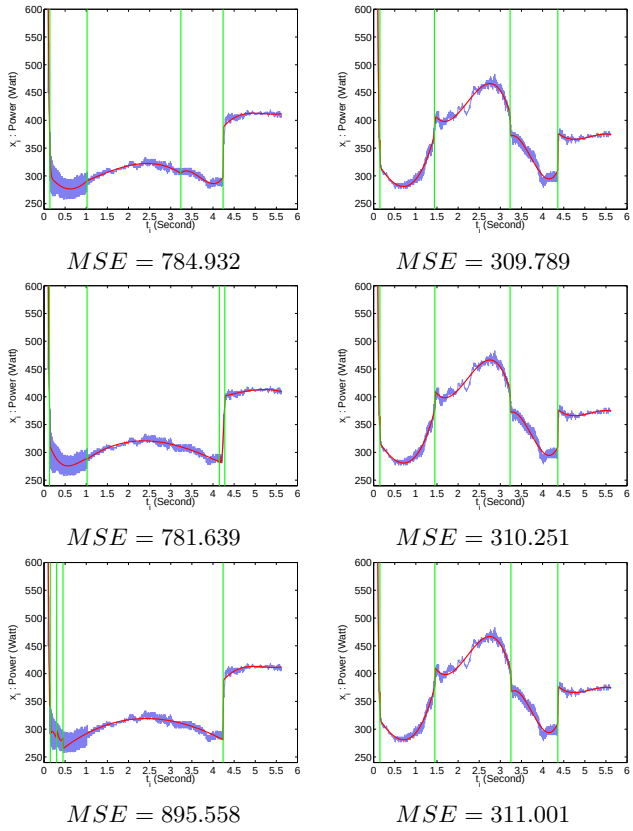


Fig. 9. Results obtained by the proposed algorithm (top), Fisher's algorithm (middle) and the iterative version of Fisher's algorithm (bottom) with the estimated model of the signal (in red), the estimated transition points (in green) and the MSE between the original signal and the estimated model.

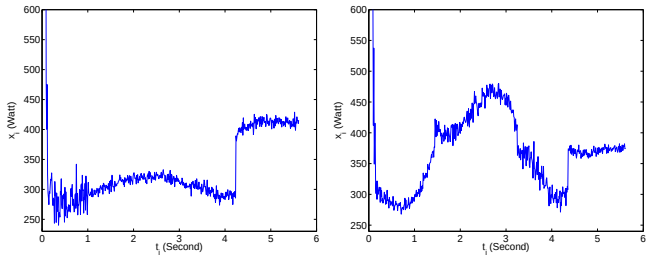


Fig. 10. Examples of generated signals corresponding to the two considered switch operations.

[5] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*, Wiley series in probability and statistics, New York, 1997.

[6] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, 6: 461-464, 1978.

[7] K. Chen, L. Xu and H. Chi, "Improved learning algorithms for Mixture of Experts in multiclass classification," *IEEE Transactions on Neural Networks*, 12(9): 1229-1252, November 1999.

[8] L.E. Baum, T. Petrie, G. Soules and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, 41: 164-171, 1970.

[9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, 77(2): 257-286, February 1989.

[10] M. Fridman, Hidden Markov Model Regression, Technical Report, Institute of mathematics, University of Minnesota, December 1993.

[11] M. I. Jordan and R. A. Jacobs, "Hierarchical Mixtures of Experts and the EM algorithm," *Neural Computation*, 6: 181-214, 1994.

[12] P. Green, "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, 1984.

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

[14] V. E. McGee and W. T. Carleton, "Piecewise regression," *Journal of the American Statistical Association*, 65, 1109-1124, 1970.

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

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

[17] V. L. Brailovsky and Y. Kempner, "Application of piece-wise regression to detecting internal structure of signal," *Pattern recognition*, 25(11), 1361-1370, November 1992.

[18] G. Ferrari-Trecate and M. Muselli, "A new learning method for piecewise linear regression," *International Conference on Artificial Neural Networks (ICANN)*, 28-30, Madrid, Spain, August 2002.

[19] A. Samé, P. Aknin and G. Govaert, "Classification automatique pour la segmentation des signaux unidimensionnels," Rencontres de la Société Francophone de Classification, ENST, Paris, 2007.