PLS approach for clusterwise linear regression on functional data

Cristian Preda[1] and Gilbert Saporta[2]

[1] Département de Statistique - CERIM, Faculté de Médecine, Université de Lille 2, 1, Place de Verdun, 59045 Lille Cedex - France cpreda@univ-lille2.fr
[2] CNAM Paris, Chaire de Statistique Appliquée, CEDRIC, 292, Rue Saint Martin, 75141 Paris Cedex 03 - France saporta@cnam.fr

Summary. Partial Least Squares approach is used for the clusterwise linear regression algorithm when the set of predictor variables forms a L_2-continuous stochastic process. The number of clusters is treated as unknown and the convergence of the clusterwise algorithm is discussed. The approach is compared with other methods via an application on stock-exchange data.

Key words: Clusterwise regression, PLS regression, principal component analysis, stochastic process.

1 Introduction

Let us consider the following regression problem : the response variable Y is numerical and the predictor set is a stochastic process $\mathbf{X} = (X_t)_{t\in[0,T]}$, $T > 0$. For example, $(X_t)_{t\in[0,T]}$ can represent temperature curves observed during a period of time $[0, T]$ and Y the amount of crops. There is here an infinite number of predictors for Y. The Partial Least Squares (PLS) approach (Wold et al. (1984)) gives better results than other tools for linear regression on a stochastic process (Preda and Saporta (2002)). In this paper we propose an improvement and extension of this technique by combining PLS regression on a stochastic process and clusterwise regression.

Clusterwise linear regression supposes that the points of each cluster are generated according to some linear regression relation : given a dataset $\{(x_j, y_j)\}_{j=1,\dots,n}$ the aim is to find simultaneously an optimal partition $\mathcal{G}$ of data in K clusters, $1 \leq K < n$, and regression coefficients $(\alpha, \beta) = \{(\alpha^i, \beta^i)\}_{i=1,\dots,K}$ within each cluster which maximize the overall fit. We write that $j \in \{1, \dots, n\}$ is an element of the cluster $i \in \{1, \dots, K\}$ by $\mathcal{G}(j) = i$. For each element j of the cluster i we consider the linear model $y_j = \alpha^i + \langle \beta^i, x_j \rangle + \epsilon_{ij}$. Then, the optimization problem associated to the clusterwise linear regression is :

$$(\mathcal{G},(\alpha,\beta)) = \operatorname{argmin} \sum_{i=1}^{K} \sum_{j:\mathcal{G}(j)=i} \epsilon_{ij}^2$$

In such a model, the parameters that have to be estimated are : the number of clusters (K), the regression coefficients for each cluster $\{(\alpha^i,\beta^i)\}_{i=1,\ldots,K}$ and the variance of residuals ϵ_{ij} within each cluster. Charles (1977) and Spaeth (1979) propose methods for estimating these parameters considering a kind of piecewise linear regression based on the least square algorithm of Bock (1969). The algorithm is a special case of k-means clustering with a criterion based on the minimization of the squared residuals instead of the classical within dispersion. The estimation of local models $\{(\alpha^i,\beta^i\}_{i=1,\ldots,K}$ could be a difficult task (number of observations less than the number of explanatory variables, multicollinearity). Solutions such as clusterwise principal component regression (PCR) or ridge regression are considered in Charles (1977). Other approaches based on mixture of distributions are developed in DeSarbo and Cron (1988) and Hennig (1999).

We propose to use Partial Least Squares (PLS) estimators (Wold et al. (1984)) for regression coefficients of each cluster in the particular case where the set of explanatory variables forms a stochastic process $\mathbf{X} = (X_t)_{t\in[0,T]}$, $T > 0$. The paper is divided into three parts. Clusterwise linear regression algorithm is presented in the first part. In the second part we recall some tools for linear regression on a stochastic process (PCR, PLS) and justify the choice for the PLS approach. The clusterwise linear regression algorithm adapted to PLS regression as well as aspects related to the prediction problem are given here and represent the main results of this work. In the last part we present an application of the clusterwise PLS regression to stock-exchange data and compare the results with those obtained by other methods such as Aguilera et al. (1998) and Preda and Saporta (2002).

2 Clusterwise linear regression

Let $\mathbf{X} = (X_j)_{j=1,\ldots,p}$, $p \geq 1$, be a real random vector and Y be a real random variable defined on the same probability space $(\Omega,\mathcal{A},P)$. We assume that $\mathbf{X}$ and Y are of second order.

The clusterwise linear regression of Y on $\mathbf{X}$ states that there exists a random variable $\mathcal{G}$, $\mathcal{G} : \Omega \to \{1,2,\ldots,K\}$, $K \in \mathbf{N} - \{0\}$ such that :

$$\begin{aligned} E(Y\big|\mathbf{X} = x, \mathcal{G} = i) &= \alpha^i + \beta^i x, \\ V(Y\big|\mathbf{X} = x, \mathcal{G} = i) &= \sigma_i^2 > 0, \forall x \in \mathbf{R}^p, \forall i = 1,\ldots K, \end{aligned} \tag{1}$$

where E and V stand for expectation, respectively variance and $\left(\alpha^i,\beta^i\right)$ are the regression coefficients associated to the cluster i, $i = 1,\ldots,K$.

Let us denote by $\hat{Y}$ the approximation of Y given by the global linear regression of Y on $\mathbf{X}$, by $\hat{Y}^i = \alpha^i + \beta^i\mathbf{X}$ the approximation of Y given by the linear regression of Y on $\mathbf{X}$ within the cluster i, $i = 1, \ldots, K$, and by $\hat{Y}^L = \sum_{i=1}^{K} \hat{Y}^i \mathbf{1}_{\mathcal{G}=i}$, where $\mathbf{1}_{\mathcal{G}=i}(j) = 1$ if j belongs to the cluster i and 0 else (L stands for "Local"). Then, the following decomposition formula holds (Charles (1977)) :

$$\begin{aligned} \mathrm{V}(Y - \hat{Y}) &= \mathrm{V}(Y - \hat{Y}^L) + \mathrm{V}(\hat{Y}^L - \hat{Y}) \\ &= \sum_{i=1}^{s} \mathbf{P}(\{\mathcal{G} = i\}) \mathrm{V}(Y - \hat{Y}^i | \mathcal{G} = i) + \mathrm{V}(\hat{Y}^L - \hat{Y}). \end{aligned} \tag{2}$$

Thus, the residual variance of the global regression is decomposed in a part of residual variance due to linear regression within each cluster and another part representing the distance between predictions given by global and local models. This formula defines the criterion for estimating the local models used in the algorithms of Charles (1977) and Bock (1969).

2.1 Estimation

Let us consider K fixed and that the homoscedasticity hypothesis holds, i.e. $\sigma_i^2 = \sigma^2$, $\forall i = 1, \ldots K$. Charles (1977) and Bock (1969) use the following criterion for estimating $\mathcal{L}(\mathcal{G})$ and $\{\alpha^i, \beta^i\}_{i=1}^K$:

$$\min_{\{\alpha^i, \beta^i\}_{i=1}^K, \mathcal{L}(\mathcal{G})} \left\{ \mathrm{V}(Y - \hat{Y}^L) \right\}. \tag{3}$$

If n data points $\{x_i, y_i\}_{i=1}^n$ have been collected, the cluster linear regression algorithm finds simultaneously an optimal partition of the n points, $\hat{\mathcal{G}}$ (as estimation of $\mathcal{L}(\mathcal{G})$), and the regression models for each cluster (element of partition) $(\hat{\alpha}, \hat{\beta} = \{\hat{\alpha}^i, \hat{\beta}^i\}_{i=1}^K$, which minimize the criterion :

$$\mathcal{V}(K, \hat{\mathcal{G}}, \hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{K} \sum_{\hat{\mathcal{G}}(j)=i} \left(y_j - (\hat{\alpha}^i + \hat{\beta}^i x_j) \right)^2. \tag{4}$$

Notice that under classical hypothesis on the model (residuals within each cluster are considered independent and normaly distributed) this criterion is equivalent to maximization of the likelihood function (Hennig (2000)).

In order to minimize (4), the clusterwise linear regression algorithms iterates the following two steps :

i) For given $\hat{\mathcal{G}}$, $\mathcal{V}(K, \hat{\mathcal{G}}, \hat{\alpha}, \hat{\beta})$ is minimezed by the LS-estimator $(\hat{\alpha}^i, \hat{\beta}^i)$ from the points (x_j, y_j) with $\hat{\mathcal{G}}(j) = i$.

ii) For given $\{\hat{\alpha}^i, \hat{\beta}^i\}_{i=1}^K$, $\mathcal{V}(K, \hat{\mathcal{G}}, \hat{\alpha}, \hat{\beta})$ is minimized according to

$$\hat{\mathcal{G}}(j) = \arg \min_{i \in \{1, \ldots, K\}} \left(y_j - (\hat{\alpha}^i + \hat{\beta}^i x_j) \right)^2. \tag{5}$$

That is, $\mathcal{V}(K, \hat{\mathcal{G}}, \hat{\alpha}, \hat{\beta})$ is monotonely decreasing if the steps $i)$ and $ii)$ are carried out alternately :

$$\underbrace{\hat{\mathcal{G}}_0 \Rightarrow (\hat{\alpha}_0, \hat{\beta}_0)}_{\mathcal{V}_0} \underset{\geq}{\Rightarrow} \underbrace{\hat{\mathcal{G}}_1 \Rightarrow (\hat{\alpha}_1, \hat{\beta}_1)}_{\mathcal{V}_1} \underset{\geq}{\Rightarrow} \ldots \underset{\geq}{\Rightarrow} \underbrace{\hat{\mathcal{G}}_l \Rightarrow (\hat{\alpha}_l, \hat{\beta}_l)}_{\mathcal{V}_l} \underset{\geq}{\Rightarrow} \ldots \tag{6}$$

where the index of each quantity denotes the iteration number, $\hat{\mathcal{G}}_0$ being an initial partition of the n data points.

In the following section we are interested to estimate the regression coefficients of each cluster in the particular case where the set of explanatory variables forms a stochastic process $\mathbf{X} = (X_t)_{t\in[0,T]}$, $T > 0$.

3 Functional data. The PLS approach

Let $\mathbf{X} = (X_t)_{t\in[0,T]}$ be a random process and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)$, $p \geq 1$, a random vector defined on the same probability space $(\Omega, \mathcal{A}, P)$. We assume that $(X_t)_{t\in[0,T]}$ and $\mathbf{Y}$ are of second order, $(X_t)_{t\in[0,T]}$ is L_2-continuous and for any $\omega \in \Omega$, $t \mapsto X_t(\omega)$ is an element of $L_2([0,T])$. Without loss of generality we assume also that $E(X_t) = 0$, $\forall t \in [0,T]$ and $E(Y_i) = 0$, $\forall i = 1, \ldots, p$.

3.1 Some tools for linear regression on functional data

It is well known that the approximation of $\mathbf{Y}$ obtained by the classical linear regression on $(X_t)_{t\in[0,T]}$, $\hat{\mathbf{Y}} = \int_0^T \beta(t) X_t dt$ is such that β is in general a distribution rather than a function of $L_2([0,T])$ (Saporta (1981)). This difficulty appears also in practice when one tries to estimate the regression coefficients, $\beta(t)$, using a sample of size N. Indeed, if $\{(\mathbf{Y}_1, \mathbf{X}_1), (\mathbf{Y}_2, \mathbf{X}_2), \ldots (\mathbf{Y}_N, \mathbf{X}_N)\}$ is a finite sample of $(\mathbf{Y}, \mathbf{X})$, the system

$$\mathbf{Y}_i = \int_0^T X_i(t)\beta(t)dt, \quad \forall i = 1, ..., N,$$

has an infinite number of solutions (Ramsay and Silverman (1997)). Regression on principal components (PCR) of $(X_t)_{t\in[0,T]}$ (Deville (1978)) and PLS approach (Preda and Saporta (2002)) give satisfactory solutions to this problem.

Linear regression on principal components (PCR)

Also known as Karhunen-Loève expansion, the principal component analysis (PCA) of the stochastic process $(X_t)_{t\in[0,T]}$ consists in representing X_t as :

$$X_t = \sum_{i\geq 1} f_i(t)\xi_i, \quad \forall t \in [0,T], \tag{7}$$

where the set $\{f_i\}_{i\geq 1}$ (the principal factors) forms an orthonormal system of deterministic functions of $L_2([0,T])$ and $\{\xi_i\}_{i\geq 1}$ (principal components) are uncorrelated zero-mean random variables (Deville (1974)). The principal factors $\{f_i\}_{i\geq 1}$ are solution of the eigenvalue equation :

$$\int_0^T C(t,s)f_i(s)ds = \lambda_i f_i(t), \tag{8}$$

where $C(t,s) = \mathrm{cov}(X_t, X_s)$, $\forall t, s \in [0,T]$. Therefore, the principal components $\{\xi_i\}_{i\geq 1}$ defined as $\xi_i = \int_0^T f_i(t)X_t dt$ are eigenvectors of the Escoufier operator, $\mathbf{W}^X$, defined by

$$\mathbf{W}^X Z = \int_0^T E(X_t Z)X_t dt, \quad Z \in L_2(\Omega). \tag{9}$$

The process $(X_t)_{t\in[0,T]}$ and the set of its principal components, $\{\xi_k\}_{k\geq 1}$, span the same linear space. Thus, the regression of $\mathbf{Y}$ on $(X_t)_{t\in[0,T]}$ is equivalent to the regression on $\{\xi_k\}_{k\geq 1}$ and we have $\hat{\mathbf{Y}} = \sum\limits_{k\geq 1} \frac{E(\mathbf{Y}\xi_k)}{\lambda_k}\xi_k$.

In practice we need to choose an approximation of order q, $q \geq 1$:

$$\hat{\mathbf{Y}}_{PCR(q)} = \sum_{k=1}^{q} \frac{E(\mathbf{Y}\xi_k)}{\lambda_k}\xi_k = \int_0^T \hat{\beta}_{PCR(q)}(t)X_t dt. \tag{10}$$

But the use of principal components for prediction is heuristic because they are computed independently of the response. The difficulty to choose the principal components used for regression is discussed in detail in Saporta (1981).

PLS regression on a stochastic process

The PLS (Partial Least Squares) approach offers a good alternative to the PCR method by replacing the least squares criterion with that of maximal covariance between $(X_t)_{t\in[0,T]}$ and $\mathbf{Y}$ (Preda and Saporta (2002)).

The PLS regression is an iterative method. Let $X_{0,t} = X_t$, $\forall t \in [0,T]$ and $\mathbf{Y}_0 = \mathbf{Y}$. At step q, $q \geq 1$, of the PLS regression of $\mathbf{Y}$ on $(X_t)_{t\in[0,T]}$, we define the q^{th} PLS component, t_q, by the eigenvector associated to the largest eigenvalue of the operator $\mathbf{W}^X_{q-1}\mathbf{W}^Y_{q-1}$, where $\mathbf{W}^X_{q-1}$, respectively $\mathbf{W}^Y_{q-1}$, are the Escoufier's operators associated to $(X_{q-1,t})_{t\in[0,T]}$, respectively to $\mathbf{Y}_{q-1}$. The PLS step is completed by the ordinary linear regression of $X_{q-1,t}$ and $\mathbf{Y}_{q-1}$ on t_q. Let $X_{q,t}$, $t \in [0,T]$ and $\mathbf{Y}_q$ be the random variables which represent the error of these regressions : $X_{q,t} = X_{q-1,t} - p_q(t)t_q$ and $\mathbf{Y}_q = \mathbf{Y}_{q-1} - \mathbf{c}_q t_q$.

Then, for each $q \geq 1$, $\{t_q\}_{q\geq 1}$ forms an orthogonal system in $L_2(X)$ and the following decomposition formulas hold :

$$\mathbf{Y} = \mathbf{c}_1 t_1 + \mathbf{c}_2 t_2 + \ldots + \mathbf{c}_q t_q + \mathbf{Y}_q,$$

$$X_t = p_1(t)t_1 + p_2(t)t_2 + \ldots + p_q(t)t_q + X_{q,t}, \quad t \in [0,T].$$

The PLS approximation of $\mathbf{Y}$ by $(X_t)_{t\in[0,T]}$ at step q, $q \geq 1$, is given by :

$$\hat{\mathbf{Y}}_{PLS(q)} = \mathbf{c}_1 t_1 + \ldots + \mathbf{c}_q t_q = \int_0^T \hat{\beta}_{PLS(q)}(t) X_t dt. \tag{11}$$

de Jong (1993) shows that for a fixed q, the PLS regression fits closer than PCR, that is,

$$R^2(\mathbf{Y}, \hat{\mathbf{Y}}_{PCR(q)}) \leq R^2(\mathbf{Y}, \hat{\mathbf{Y}}_{PLS(q)}). \tag{12}$$

In Preda and Saporta (2002)) we show the convergence of the PLS approximation to the approximation given by the classical linear regression :

$$\lim_{q\to\infty} E(\|\hat{\mathbf{Y}}_{PLS(q)} - \hat{\mathbf{Y}}\|^2) = 0. \tag{13}$$

In practice, the number of PLS components used for regression is determined by cross-validation (Tenenhaus (1998)).

Remark 1 (Numerical solution). Because $(X_t)_{t\in[0,T]}$ is a continous time stochastic process, in practice we need a discretization of the time interval in order to obtain a numerical solution. In Preda (1999) we give such an approximation. Thus, if $\Delta = \{0 = t_0 < t_1 < \ldots < t_p = T\}$, $p \geq 1$, is a discretization of $[0,T]$, consider the process $(X_t^{\Delta})_{t\in[0,T]}$ defined as :

$$X_t^{\Delta} = \frac{1}{t_{i+1} - t_i} \int_{t_i}^{t_{i+1}} X_t dt, \quad \forall t \in [t_i, t_{i+1}[, \quad \forall i = 0, ..., p-1. \tag{14}$$

Denote by m_i the random variable $\frac{1}{t_{i+1}-t_i}\int_{t_i}^{t_{i+1}} X_t dt$, $i = 0, \ldots, p-1$ which represents the time-average of X_t on the interval $[t_i, t_{i+1}[$. Then, the approximation of the PLS regression on $(X_t)_{t\in[0,T]}$ by that on $(X_t^{\Delta})_{t\in[0,T]}$ is equivalent to the PLS regression on the finite set $\{m_i\sqrt{t_{i+1} - t_i}\}_{i=0,\ldots,p-1}$.

For some fixed p, we give in Preda (1999) a criterion for the choice of the optimal discretization Δ according to the approximation given in (14).

3.2 The clusterwise PLS approach

When the predictor variables form a stochastic process $\mathbf{X} = (X_t)_{t\in[0,T]}$, i.e. x_i are curves, the classical linear regression is not adequate to give estimators for the linear models within clusters, $\{\alpha^i, \beta^i\}_{i=1}^K$ (Preda and Saporta (2002)). We propose to adapt the PLS regression for the clusterwise algorithm in order to overcome this problem. Thus, the local models are estimated using the PLS approach given in the previous section. Notice that PCR and Ridge Regression approaches are discussed for the discrete case (finit number of predictor variables) in Charles (1977).

Let us denote by $\hat{\mathcal{G}}_{PLS,s}$, $\{\hat{\alpha}^i_{PLS,s}, \hat{\beta}^i_{PLS,s}\}_{i=1}^K$ this estimators at the step s of the clusterwise algorithm using the PLS regression. However, a natural question arises : is the clusterwise algorithm still convergent in this case ? Indeed, the LS criterion is essential in the proof of the convergence of the algorithm (Charles (1977)). The following proposition justify the use of the PLS approach in this context :

Proposition 1. *For each step s of the clusterwise PLS regression algorithm there exists a pozitive integer $q(s)$ such that $\hat{\mathcal{G}}_{PLS,s}$ and $\{\hat{\alpha}^i_{PLS,s}, \hat{\beta}^i_{PLS,s}\}_{i=1}^K$ given by the PLS regressions using $q(s)$ PLS components preserve the decreasing monotonicity of the sequence $\{\mathcal{V}(\hat{\mathcal{G}}_{PLS,s},\ \{\hat{\alpha}^i_{PLS,s}, \hat{\beta}^i_{PLS,s}\}_{i=1}^K)\}_{s\geq 1}$*

Proof. Let $s \geq 1$ and $\left(\hat{\mathcal{G}}_{PLS,s}, \{\hat{\alpha}^i_{PLS,s}, \hat{\beta}^i_{PLS,s}\}_{i=1}^K\right)$ be the estimators given by the clusterwise PLS algorithm at the step s. From (5) we have on one hand

$$\mathcal{V}(\hat{\mathcal{G}}_{PLS,s}, \{\alpha^i_{PLS,s}, \beta^i_{PLS,s}\}_{i=1}^K) \geq \mathcal{V}(\hat{\mathcal{G}}_{PLS,s+1}, \{\hat{\alpha}^i_{PLS,s}, \hat{\beta}^i_{PLS,s}\}_{i=1}^K).$$

On the other hand, from (13), there exists $q(s+1)$ such that :

$$\mathcal{V}(\hat{\mathcal{G}}_{PLS,s+1}, \{\alpha^i_{PLS,s}, \beta^i_{PLS,s}\}_{i=1}^K) \geq \mathcal{V}(\hat{\mathcal{G}}_{PLS,s+1}, \{\hat{\alpha}^i_{PLS,s+1}, \hat{\beta}^i_{PLS,s+1}\}_{i=1}^K),$$

where $\{\hat{\alpha}^i_{PLS,s+1}, \hat{\beta}^i_{PLS,s+1}\}_{i=1}^K$ are the estimators of the regression coeficients within each cluster using $q(s+1)$ PLS components. Thus the proof is complete.

Remark 2. In our experiences, the number of PLS components given by cross-validation with thereshold (Tenenhaus (1998)) higher than 0.9 provides a good approximation for $q(s)$ from Proposition 1.

Let us denote by $\hat{\mathcal{G}}_{PLS}$, $\{\hat{\alpha}^i_{PLS}, \hat{\beta}^i_{PLS}\}_{i=1}^K$ the estimators given by the clusterwise algorithm using the PLS regression.

Prediction

Given a new data point (x_{i^*}, y_{i^*}) for which we have only the observation of x_{i^*}, the prediction problem of y_{i^*} is reduced to the problem of pattern recognition : to which cluster belongs the point (x_{i^*}, y_{i^*}) ? A rule that uses the k-nearest neighbours approach is proposed by Charles (1977) :
Let m, M be two positive integers such that $m \leq M$. For each $k \in [m, M]$ let

- N_k be the set of the k nearest neighbours of x_{i^*},
- $n_j(k) = \left|G^{-1}(j) \cap N_k\right|$, $\forall j \in 1, \ldots, K$,
- $J(k) = \{j \in \{1, \ldots, K\} : n_j(k) = \max\limits_{l=1,\ldots,K} n_l(k)\}$,
- $\eta_j = \sum_{k=m}^M \mathbf{1}_{J(k)}(j)$

Then,

$$\mathcal{G}(i^*) = \arg\max_j \eta_j. \tag{15}$$

Therefore,

$$\hat{y}_{i^*} = \hat{\alpha}_{PLS}^{\hat{\mathcal{G}}(i^*)} + \int_0^T \hat{\beta}_{PLS}^{\hat{\mathcal{G}}(i^*)}(t) x_{i^*}(t) dt. \qquad (16)$$

It is important to notice that the properties of the clusterwise PLS regression do not change if $\mathbf{Y}$ is a random vector of finite or infinite dimension. When $\mathbf{Y} = \{X_t\}_{t\in[T,T+a]}$, the clusterwise PLS regression is used to predict the future of the process from its past. We will consider this situation in the following application on stock exchange data. The number of clusters is considered unknown. Charles (1977) proposes to choose K observing the behavior of the decreasing function $c(K) = \frac{V(Y-\hat{Y}^L)}{V(Y)}$. Other criteria based on the decomposition formula (2) are proposed in Plaia (2001). Hennig (1999) gives some approximations of K based on the likelihood function.

4 Application on stock exchange data

The clusterwise PLS regression on a stochastic process presented in the previous sections is used to predict the behavior of shares on a certain lapse of time. We have developed a C++ application which implements the clusterwise PLS approach, by varying the number of clusters and using the cross-validation.

We have 84 shares quoted at the Paris stock exchange, for which we know the whole behavior of the growth index during one hour (between 10^{00} and 11^{00}). Notice that a share is likely to change every second. We also know the evolution of the growth index of a new share (quoted 85) between 10^{00} and 10^{55}. The aim is to predict the way that share will behave between 10^{55} and 11^{00} using the clusterwise PLS approach built with the other 84 shares.

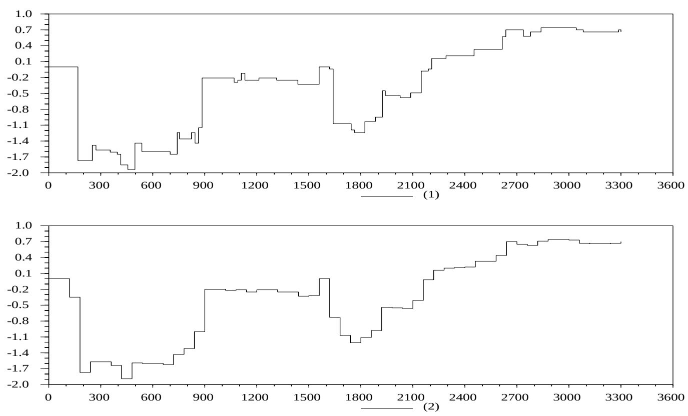

Fig. 1. Evolution of the share 85 : (1) – before approximation (2) – after approximation

We use the approximation given in (14) by taking an equidistant discretization of the interval $[0, 3600]$ (time expressed in seconds) in 60 subintervals. Figure 1 gives the evolution of the share 85 in $[0, 3300]$ before and after this approximation. The forecasts obtained will then match the average level of the growth index of share 85 considered on each interval $[60 \cdot (i-1), 60 \cdot i)$, $i = 56, \ldots, 60$.

The same data are used in Preda and Saporta (2002) where the global PCR and PLS regressions are fitted. We quote by CW-PLS(k) the clusterwise PLS regression with k clusters, by PCR(k) respectively PLS(k), the global regression on the first k principal components, respectively on the first k PLS components.

Considering the cross-validation approach (thereshold = 0.95) we obtain the following results :

	$\hat{m}_{56}(85)$	$\hat{m}_{57}(85)$	$\hat{m}_{58}(85)$	$\hat{m}_{59}(85)$	$\hat{m}_{60}(85)$	SSE
Observed	**0.700**	**0.678**	**0.659**	**0.516**	**-0.233**	-
PLS(2)	0.312	0.355	0.377	0.456	0.534	0.911
PLS(3)	0.620	0.637	0.677	0.781	0.880	1.295
PCR(3)	0.613	0.638	0.669	0.825	0.963	1.511
CW-PLS(3)	0.643	0.667	0.675	0.482	0.235	0.215
CW-PLS(4)	0.653	0.723	0.554	0.652	-0.324	0.044
CW-PLS(5)	0.723	0.685	0.687	0.431	-0.438	0.055

Using the sum of squared errors (SSE) as measure of fit, let us observe that the clusterwise models give better results than the global analysis. The models with 4 and 5 clusters predict the crash of the share 85 for the last 5 minutes, whereas the global models do not. For the model with 4 clusters, the estimation of the distribution of $\mathcal{G}$ is $(\frac{17}{84}, \frac{32}{84}, \frac{10}{84}, \frac{25}{84})$, the point 85 belonging to the first cluster.

5 Conclusions

The clusterwise PLS regression on a stochastic process offers an interesting alternative to classical methods of clusterwise analysis. It is particularly adapted to solve multicollinearity problems for regression and also when the number of observations is smaller than the number of predictor variables, which is often the case in the context of the clusterwise linear regression.

References

Aguilera, A. M., Ocaña, F. and Valderrama, M. J. (1998). An approximated principal component prediction model for continuous-time stochastic process. *Applied Stochastic Models and Data Analysis* **11** 61–72.

Bock, H. H. (1969). The equivalence of two extremal problems and its application to the iterative classification of multivariate data. *Lecture note, Mathematisches Forschungsinstitut Oberwolfach* .

Charles, C. (1977). Régression typologique et reconnaissance des formes. Thèse de doctorat. Tech. Rep., Université Paris IX, Université Paris IX, Paris, France.

DeSarbo, W. S. and Cron, W. L. (1988). A Maximum Likelihood Methodology for Clusterwise Linear Regression. *Journal of Classification* **5** 249–282.

Deville, J. C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE (France)* 3–101.

Deville, J. C. (1978). Analyse et prévision des séries chronologiques multiples non stationnaires. *Statistique et Analyse des Données (France)* **3** 19–29.

Hennig, C. (1999). Models and Methods for Clusterwise Linear Regression. *Classification in the Information Age, Springer, Berlin* 179–187.

Hennig, C. (2000). Identifiability of Models for Clusterwise Linear Regression. *Journal of Classification* **17** 273–296.

Plaia, A. (2001). On the number of clusters in clusterwise linear regression. *X-th International Symposium on Applied Stochastic Models and data analysis, Proceedings* 847–852.

Preda, C. (1999). Analyse factorielle d'un processus : problèmes d'approximation et de régression. Thèse de doctorat. Tech. Rep. 2648, Université de Lille 1, Université de Lille 1, Lille, France.

Preda, C. and Saporta, G. (2002). Régression PLS sur un processus stochastique. *Revue de Statistique Appliquée* **50** 27–45.

Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis, Springer Series in Statistics*. Springer-Verlag, New York.

Saporta, G. (1981). Méthodes exploratoires d'analyse de données temporelles, Cahiers du B.U.R.O. Tech. Rep. 37, Université Pierre et Marie Curie, Université Pierre et Marie Curie, Paris, France.

Spaeth, H. (1979). Clusterwise linear regression. *Computing* **22** 367–373.

Tenenhaus, M. (1998). *La régression PLS. Théorie et pratique*. Editions Technip, Paris.

Wold, S., Ruhe, A. and Dunn III, W. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **5** 735–743.