Clusterwise PLS regression on a stochastic process

Cristian Preda[1] and Gilbert Saporta[2]

[1] CERIM - Département de Statistique, Faculté de Médecine, Université de Lille 2, 59045 Lille Cedex, France
[2] Chaire de Statistique Appliquée, Conservatoire National des Arts et Métiers, Paris Cedex 03, France

Abstract. In this paper we propose to use the PLS approach for clusterwise linear regression in the particular case where the set of predictor variables forms a L_2-continuous stochastic process $\{X_t\}_{t\in[0,T]}$. We have adapted the k-means algorithm to this case and we give necessar conditions for its convergence. The results of an application of the clusterwise PLS regression to stock-exchange data are compared with those obtained by other methods.

Keywords. Clusterwise regression, PLS regression, stochastic process

1 Introduction

According to Charles (1977) the clusterwise linear regression is defined as a kind of piecewise regression : given a data set $\{(X_i, Y_i)\}_{i=1}^n$ of observations of an explanatory variable X and a response variable Y, the aim is to find simultaneously an optimal partition of the data and the regression models associated to each cluster which maximize the overall fit.

When the regression is used for prediction such an approach gives usually better results than a global regression. The algorithm is a special case of k-means clustering with a criterium based on the minimisation of the squared residuals instead of the classical within class dispersion. However the estimation of the local models could be a difficult task (number of observations less than the number of variables, multicollinearity, etc). Solutions such as local PCR regression or local ridge regression may solve these difficulties (Charles, 1977). Algorithms for the least squares solution are also given in Spaeth (1979).

In this paper we propose to use PLS regression (Wold et al., 1994) for each cluster in the particular case where the set of predictors forms a stochastic process $X = \{X_t\}_{t\in[0,T]}$. In other words the problem consists in predicting a Y variable by a set of curves. Thus, clusterwise PLS regression on a stochastic process is an extension of the global PLS approach given in Preda and Saporta (2001).

The paper is divided into three parts. In the first part we introduce some tools for linear regression on a stochastic process (PCR, PLS) and justify the choice of the PLS approach. The clusterwise linear regression algorithm adapted to the case of PLS regression is discussed in the second part of the paper. Aspects related to the prediction problem are also presented. 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 (2001).

2 Some tools for linear regression on a stochastic process

Let $(X_t)_{t\in[0,T]}$ be a random process and Y a random variable defined on the same probability space $(\Omega, \mathcal{A}, P)$. We assume that $(X_t)_{t\in[0,T]}$ and 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 $E(Y_i) = 0$, $\forall i = 1, \ldots, p$.

It is well known that the linear model obtained by the classical regression of Y on $(X_t)_{t\in[0,T]}$, $\hat{Y} = \int_0^T \beta(t)X_t dt$, is such that β is in general a distribution instead a function of $L_2([0,T])$ (Saporta, 1981). Regression on principal components of $(X_t)_{t\in[0,T]}$ and PLS regression give satisfactory solution to this problem.

2.1 Linear regression on principal components

The principal components regression (PCR) regress the responses onto the principal components of the set of explanatory variables. The principal components of the stochastic process $\{X_t\}_{t\in[0,T]}$ is the family $\{\xi_i\}_{i\geq 1}$ of uncorrelated zero-mean random variables defined by $\xi_i = \int_0^T f_i(t)X_t dt$, where $\{f_i\}_{i\geq 1}$ (the principal factors) is the orthonormal family of eigenfunctions of covariance operator of the process $\{X_t\}_{t\in[0,T]}$ associated to its decreasing sequence of non null eigenvalues $\{\lambda_i\}_{i\geq 1}$. The principal components are also eigenvectors of the Escoufier operator, $\mathbf{W}^X$, defined by $\mathbf{W}^X Z = \int_0^T E(X_t Z)X_t dt$, $Z \in L_2(\Omega)$. Therefore, the process can be represented as $X_t = \sum_{i\geq 1} \xi_i f_i(t)$, $\forall t \in [0,T]$. A such representation is called the Karhunen-Loève expansion of the process (Saporta, 1981).

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 Y on $(X_t)_{t\in[0,T]}$ is equivalent to the regression on $\{\xi_k\}_{k\geq 1}$ and we have $\hat{Y} = \sum_{k\geq 1} \frac{E(Y\xi_k)}{\lambda_k}\xi_k$.
In practice we need to choose an approximation of order q, $q \geq 1$:

$$\hat{Y}^q_{PCR} = \sum_{k=1}^{q} \frac{E(Y\xi_k)}{\lambda_k}\xi_k = \int_0^T \hat{\beta}_{PCR}(t)X_t dt. \tag{2.1}$$

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

2.2 PLS regression on a stochastic proces

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 Y (Preda and Saporta, 2001).

The PLS regression is an iterative method. Let $X_{0,t} = X_t$, $\forall t \in [0,T]$ and $Y_0 = Y$. At the step q, $q \geq 1$, of the PLS regression of 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 Y_{q-1}. The PLS step is completed by the ordinary linear regression of $X_{q-1,t}$ and Y_{q-1} on t_q. Let $X_{q,t}$, $t \in [0,T]$ and 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 $Y_q = Y_{q-1} - c_q t_q$.

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

$$Y = c_1 t_1 + c_2 t_2 + \ldots + c_q t_q + Y_q,$$
$$X_t = p_1(t)t_1 + p_2(t)t_2 + \ldots + p_q(t)t_q + X_{q,t}, \quad t \in [0,T].$$

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

$$\hat{Y}^q_{PLS} = c_1 t_1 + \ldots + c_q t_q = \int_0^T \hat{\beta}_{PLS}(t) X_t dt. \tag{2.2}$$

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

$$R^2(Y, Y^q_{PCR}) \leq R^2(Y, Y^q_{PLS}). \tag{2.3}$$

In Preda and Saporta (2001) we show the convergence of the PLS approximation to the approximation given by the classical linear regression :

$$\lim_{q\to\infty} E(|\hat{Y}_q - \hat{Y}|^2) = 0. \tag{2.4}$$

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

3 Clusterwise PLS regression

The clusterwise linear regression supposes that there exists a group-variable $G : \Omega \to \{1, 2, \ldots, k\}$, $1 \leq k < \infty$, such that

$$E(Y|X = x, G = i) = \alpha^{(i)} + \beta^{(i)}x, \quad \forall i = 1, \ldots k,$$

where $\{\alpha^{(i)}, \beta^{(i)}\}_{i=1}^k$ are subgroup-specific regression coefficients given by the least squares criterion. Let us denote by

$\hat{Y} = \alpha + \beta X$, the approximation given by the global regression of Y on X,

$\hat{Y}^{(i)} = \alpha^{(i)} + \beta^{(i)}X, \quad \forall i = 1, \ldots k,$ and $\hat{Y}^L = \sum_{i=1}^k \hat{Y}^{(i)} \mathbf{1}_{\{G=i\}}$.

Then we have

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

In general k is unknown, and therefore, the distribution $\mathcal{L}(G)$ of G is unknown.

For fixed k, the clusterwise linear regression gives estimation of $\mathcal{L}(G)$ and $\{\alpha^{(i)}, \beta^{(i)}\}_{i=1}^{k}$ using the criterion : $\min_{\{\alpha^{(i)},\beta^{(i)}\}_{i=1}^{k},\mathcal{L}(G)} \Big\{ \mathrm{V}ar(Y - \hat{Y}^{L}) \Big\}$.

3.1 The clusterwise linear regression algorithm

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, $P = (P^{(1)}, \ldots, P^{(k)})$ (as estimation of $\mathcal{L}(G)$), and the regression models associated to each cluster, $\{\alpha^{(i)}, \beta^{(i)}\}_{i=1}^{k}$, which maximize the criterion :

$$L(P, \{\alpha^{(i)}, \beta^{(i)}\}_{i=1}^{k}) = \sum_{i=1}^{k} \sum_{j \in P^{(i)}} \Big(Y_j - (\alpha^{(i)} + \beta^{(i)} X_j) \Big)^2$$

Starting with an initial partition $P_0 = (P_0^{(1)}, \ldots, P_0^{(k)})$ the algorithm constructs iteratively a sequence $(P_s, \{\alpha_s^{(i)}, \beta_s^{(i)}\}_{i=1}^{k})_{s \geq 0}$ in the following way (Charles, 1977) :

- for each $i = 1, \ldots k$, $(\alpha_0^{(i)}, \beta_0^{(i)})$ are given by the least square estimators of the linear regression using the data points of the cluster $P_0^{(i)}$.
- let $(P_s, \{\alpha_s^{(i)}, \beta_s^{(i)})$ be known. Then, for each $i = 1, \ldots, k$,

$$P_{s+1}^{(i)} = \left\{ (Y_j, X_j) | \Big(Y_j - (\alpha_s^{(i)} + \beta_s^{(i)} X_j) \Big)^2 < \Big(Y_j - (\alpha_s^{(i')} + \beta_s^{(i')} X_j) \Big)^2 , \forall i \neq i' \right\},$$

$(\alpha_{s+1}^{(i)}, \beta_{s+1}^{(i)})$ are the least squares estimators of linear regression using the data points of the cluster $P_{s+1}^{(i)}$.

The sequence $(P_s, \{\alpha_s^{(i)}, \beta_s^{(i)}\}_{i=1}^{k})_{s \geq 0}$ is such that

$$L(P_s, \{\alpha_s^{(i)}, \beta_s^{(i)}\}_{i=1}^{k}) \geq L(P_{s+1}, \{\alpha_{s+1}^{(i)}, \beta_{s+1}^{(i)}\}_{i=1}^{k}), \forall s \geq 0$$

and so it is convergent. Therefore, $(P_s, \{\alpha_s^{(i)}, \beta_s^{(i)}\}_{i=1}^{k})_{s \geq 0}$ is convergent and reaches its limit. Let $(P, \{\hat{\alpha}^{(i)}, \hat{\beta}^{(i)}\}_{i=1}^{k})$ be this limit.

3.2 Clusterwise PLS regression when data are curves

When explanatory variables are curves, $X_i \in L_2([0, T])$, the classical linear regression is not adequate to give estimators for the local models $\{(\alpha^{(i)}, \beta^{(i)}\}_{i=1}^{k})$ (Preda and Saporta, 2001).

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. Let us denote by $\{(\alpha_{PLS,s}^{(i)}, \beta_{PLS,s}^{(i)}\}_{i=1}^{k})$ this estimators at the step s of algorithm.

However, a natural question arises : is the clusterwise algorithm still convergent in this case ? Indeed, the least squares criterion is essential in the proof of the convergence of the algorithm when the set of explanatory variables is finite (Charles, 1977). The following proposition gives the answer to

this question.

Proposition 3.1 *For each step s of the clusterwise PLS regression algorithm there exists $q(s)$, $q(s) \geq 1$, such that the local PLS regressions with $q(s)$ PLS components preserve the convergence of the algorithm.*

Proof : Let $\left(P_s, \{(\alpha^{(i)}_{PLS,s}, \beta^{(i)}_{PLS,s}\}_{i=1}^k)\right)$ be the associate estimators for each local model at the step s of the clusterwise PLS algorithm. By construction we have that

$$L(P_s, \{\alpha^{(i)}_{PLS,s}, \beta^{(i)}_{PLS,s}\}_{i=1}^k) \geq L(P_{s+1}, \{\alpha^{(i)}_{PLS,s}, \beta^{(i)}_{PLS,s}\}_{i=1}^k).$$

On the other hand, from (2.4) there exists $q(s+1)$ such that :

$$L(P_s, \{\alpha^{(i)}_{PLS,s+1}, \beta^{(i)}_{PLS,s}\}_{i=1}^k) \geq L(P_{s+1}, \{\alpha^{(i)}_{PLS,s+1}, \beta^{(i)}_{PLS,s+1}\}_{i=1}^k) \geq L(P_{s+1}, \{\alpha^{(i)}_{s+1}, \beta^{(i)}_{s+1}\}_{i=1}^k),$$

where $\{\alpha^{(i)}_{s+1}, \beta^{(i)}_{s+1}\}_{i=1}^k$ are the estimators given by least squares for each cluster at the step $s+1$. The proof is complete.

From practical point of view, that result allows the use of the cross-validation criterion (with right parameters) in order to perform clusterwise PLS regression.

Let us denote the by $\{(\hat{\alpha}^{(i)}_{PLS}, \hat{\beta}^{(i)}_{PLS}\}_{i=1}^k)$ the PLS estimators for each cluster given by the clusterwise PLS regression.

Prediction. Given a new data point (Y^*, X^*) for which one has only the observation of X, the prediction problem of Y^* is reduced to those of the determination (choice) of the cluster which contains this point. A rule that use the approach of the k-nearest neighbours is proposed by Charles (1977). If $P^{(i^*)}$ is the cluster given by this rule, then the prediction of Y^* given by the clusterwise PLS regression is

$$\hat{Y}^* = \hat{\alpha}^{(i^*)}_{PLS} + \int_0^T \hat{\beta}^{(i^*)}_{PLS}(t) X^*(t) dt.$$

It is important to notice that the properties of the clusterwise PLS regression do not change if Y is a random vector of finite or infinite dimension. This extension of the PLS regression is given in Preda and Saporta (2001). When $Y = \{X_t\}_{t \in [T, T+a]}$ the clusterwise PLS regression is used to predict the future of the process from its past.

Number of clusters. The number of clusters, k, is unknown. Charles (1977) proposed to choose k by observing the evolution of the decreasing function $c(k) = \dfrac{Var(Y - \hat{Y}^L)}{Var(Y)}$. Other criteria based on the same formula of decomposition of variance are proposed in Plaia (2001).

4 Application on stock exchange data

The clusterwise PLS regression on a stochastic process presented in the previous sections is used to predict the behaviour 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 cluster and using the cross-validation criteria for different level of significance of PLS components.

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}) ; a share is likely to change every second. We also know the evolution of the growth index of a new share (noted 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. The same data are used in Preda and Saporta (2001) where the global PCR and PLS regressions are fitted. We denote 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.

Using the same approximation and notations as in Preda and Saporta (2001) we have obtained 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

References

Aguilera, A.M., Ocaña F. and Valderrama, M.J. (1998): *An approximated principal component prediction model for continous-time stochastic process*, Applied Stochastic Models and Data Analysis, Vol. 13, p. 61-72

Charles, C. (1977) *Regression typologique et reconnaissance des formes*, These de doctorat 3eme cycle, Universite Paris IX.

Deville, J. C. (1974) : *Méthodes statistiques et numériques de l'analyse harmonique*, Annales de l'INSEE, No. 15, p 3-101.

Green, P.J. and Silverman, B. W. (1994): *Nonparametric Regression and generalized linear models. A roughness penalty approach*, Monographs on statistic and applied probability, No. 58, Chapman & Hall.

de Jong, S. (1993) : *PLS fits closer than PCR*, Journal of Chemometrics, 7, 551-557.

Palm, R. and Iemma, A.F. (1995): *Quelques alternatives a la regression classique dans le cas de colinearite*, Rev. Statistique Appliquee XLIII (2), p. 5-33.

Plaia, A. (2001) : *On the number of clusters in clusterwise linear regression*, X-th International Symposium on Applied Stochastic Models and data analyis, Proceedings, Volume 2, p. 847-852, Compiegne-France.

Preda, C., Saporta, G. (2001) : *PLS regression on a stochastic process*, X-th International Symposium on Applied Stochastic Models and data analyis, Proceedings, Volume 2, p. 853-859, Compiegne-France.

Saporta, G. (1981) : *Méthodes exploratoires d'analyse de données temporelles*, Cahiers du B.U.R.O., No. 37-38, Université Pierre et Marie Curie, Paris.

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., Dunn III, W.J. (1984) *The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses*, SIAM J. Sci. Stat. Comput., vol 5, no.3 pp. 735-743.