Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis 49 (2005) 99 – 108

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Clusterwise PLS regression on a stochastic process

C. Preda[a,*], G. Saporta[b]

[a]Département de Statistique-CERIM, Faculté de Médecine, Université de Lille 2,
1, Place de Verdun, 59045 Lille, Cedex, France
[b]CNAM Paris, Chaire de Statistique Appliquée, CEDRIC, 292, Rue Saint Martin, 75141 Paris, Cedex 03, France

Received 27 May 2003; received in revised form 4 May 2004; accepted 4 May 2004
Available online 9 June 2004

Abstract

The clusterwise linear regression is studied when the set of predictor variables forms a L_2-continuous stochastic process. For each cluster the estimators of the regression coefficients are given by partial least square regression. 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.

Keywords: Clusterwise regression; PLS regression; Principal component analysis; Stochastic process

1. Introduction

Cluster analysis based on stochastic models considers that a cluster is a subset of data points which can be modeled adequately in order to reflect the meaning of homogeneity with respect to the certain data analysis problem. The clusterwise linear regression supposes that the points of each cluster are generated according to some linear regression relation: given a dataset $\{(x_i, y_i)\}_{i=1,\dots,n}$ the aim is to find simultaneously an optimal partition $\mathscr{G}$ of data in K clusters, $1 \leqslant K < n$, and regression coefficients $(\alpha, \beta) = \{(\alpha^i, \beta^i)\}_{i=1,\dots,K}$ within

[*] Corresponding author. Tel.: +33-320-62-69-69; fax: +33-320-52-10-22.
E-mail addresses: cpreda@univ-lille2.fr (C. Preda), saporta@cnam.fr (G. Saporta).

doi:10.1016/j.csda.2004.05.002

each cluster which maximize the overall fit:

$$
\begin{aligned}
&\mathscr{G}:\{1,\ldots,n\}\rightarrow\{1,\ldots,K\},\quad \mathscr{G}^{-1}(i)\neq\phi,\quad \forall i=1,\ldots,K,\\
&\forall\, i=1,\ldots,K,\quad y_j=\alpha^i+\langle\beta^i,x_j\rangle+\varepsilon_{ij},\quad \forall j:\mathscr{G}(j)=i,\\
&(\mathscr{G},(\alpha,\beta))=\operatorname{argmin}\sum_{i=1}^{K}\sum_{j:\mathscr{G}(j)=i}\varepsilon_{ij}^2.
\end{aligned}
$$

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 (CW-PCR) or ridge regression (RR) are considered in Charles (1977). The partial least squares (PLS) approach (Wold et al., 1984) is considered for finite number of predictors by Esposito Vinzi and Lauro (2003) as PLS typological regression. Other approaches based on mixture of distributions are developed in DeSarbo and Cron (1988), Hennig (2000) and Hennig (1999).

In this paper, we propose to use the PLS estimators for regression coefficients of each cluster in the particular case where the set of explanatory variables forms a stochastic process $\mathbf{X}=(X_t)_{t\in[0,T]}$, $T>0$. Thus, clusterwise PLS regression on a stochastic process is an extension of the global PLS approach given in Preda and Saporta (2002). 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 PLS regression as well as aspects related to the prediction problem are discussed in the second part. 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. (1997) and Preda and Saporta (2002).

2. Some tools for linear regression on a stochastic process

Let $\mathbf{X}=(X_t)_{t\in[0,T]}$ be a random process and $\mathbf{Y}=(Y_1,Y_2,\ldots,Y_p)$, $p\geqslant 1$, a random vector defined on the same probability space $(\Omega,\mathscr{A},P)$. 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 $E(Y_i)=0$, $\forall i=1,\ldots,p$.

It is well known that the approximation of $\mathbf{Y}$ obtained by the classical linear regression on $(X_t)_{t\in[0,T]}$, $\hat{\mathbf{Y}}=\int_0^T\beta(t)X_t\,\mathrm{d}t$ 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