Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis 48 (2005) 149–158

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

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, 59043 Lille Cedex, France

[b]CNAM Paris, Chaire de Statistique Appliquée, CEDRIC, 292, Rue Saint Martin, 75141 Paris Cedex 03, France

Received 9 October 2003; received in revised form 9 October 2003; accepted 9 October 2003

Abstract

Partial least squares (PLS) regression on an L_2-continuous stochastic process is an extension of the finite set case of predictor variables. The PLS components existence as eigenvectors of some operator and convergence properties of the PLS approximation are proved. The results of an application to stock-exchange data will be compared with those obtained by other methods.

Keywords: PLS regression; Stochastic process; Escoufier's operator; Principal component analysis

1. Introduction

It does not seem usual to perform a linear regression when the number of predictors is infinite. However, it is the case when one tries to predict a response variable Y thanks to the observation of a time-dependent variable X_t, for any $t \in [0, T]$ (for example, $(X_t)_{t\in[0,T]}$ can represent temperature curves observed in n places and Y the amount of crops). Theoretically, this can be expressed by the regression of Y on the process $(X_t)_{t\in[0,T]}$.

The aim of this paper is to adapt the PLS regression when the set of predictor variables forms a stochastic process Fig. 1. The problems brought about by the classical linear regression on a process—the indetermination of the regression coefficients (Ramsay and Dalzell, 1991; Ramsay and Silverman, 1997; Saporta, 1981) or the choice of the principal components of $(X_t)_{t\in[0,T]}$ as predictor variables (Deville, 1978;

* Corresponding author. Tel.: +33-3-20-62-69-69; fax: +33-3-20-52-10-22.
E-mail addresses: cpreda@univ-lille2.fr (C. Preda), saporta@cnam.fr (G. Saporta).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2003.10.003

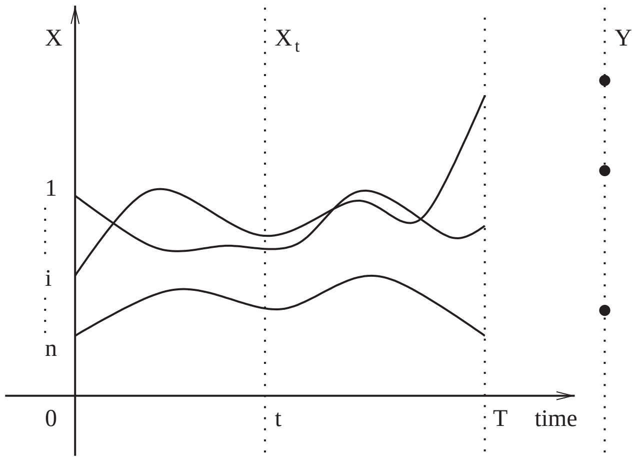

Fig. 1. Regression on a stochastic process.

Saporta, 1981; Aguilera et al., 1998)—get satisfactory solutions within this framework, the main characteristics which are derived from those of the Escoufier operator associated with the process $(X_t)_{t\in[0,T]}$.

PLS regression on a stochastic process is an extension of the finite case (finite set of predictors) developed by Wold et al. (1984), Tenenhaus et al. (1995) and Cazes (1997) (see also Eldén, 2003; Nguyen and Rocke, 2003).

We prove the existence of PLS components as well as some convergence properties towards the classical linear regression. The case $\mathbf{Y} = (X_t)_{t\in[T,T+a]}$, $a > 0$, presents an alternative to prediction methods proposed by Aguilera et al. (1998) and Deville (1978). The results of an application on stock exchange data are compared with those obtained by other methods.

2. Some results about principal component analysis (PCA) and regression when data are curves

Let $(X_t)_{t\in[0,T]}$ 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 also assume that $E(X_t) = 0$, $\forall t \in [0,T]$ and $E(Y_i) = 0$, $\forall i = 1, \ldots, p$.

2.1. PCA of a stochastic process

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:

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