Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 4832–4848
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Smoothing splines estimators in functional linear regression with errors-in-variables

Hervé Cardot[a], Christophe Crambes[b,*], Alois Kneip[c], Pascal Sarda[b]

[a] INRA Dijon, CESAER - ENESAD, 26, bd Docteur Petitjean, BP 87999, 21079 Dijon Cedex, France
[b] Laboratoire de Statistique et Probabilités, Université Paul Sabatier, UMR C5583, 118, route de Narbonne, 31062 Toulouse Cedex, France
[c] Statistische Abteilung, Department of Economics, Universität Bonn, Adenauerallee 24, 53113 Bonn, Germany

Available online 15 August 2006

Abstract

The total least squares method is generalized in the context of the functional linear model. A smoothing splines estimator of the functional coefficient of the model is first proposed without noise in the covariates and an asymptotic result for this estimator is obtained. Then, this estimator is adapted to the case where the covariates are noisy and an upper bound for the convergence speed is also derived. The estimation procedure is evaluated by means of simulations.

Keywords: Functional linear model; Smoothing splines; Penalization; Errors-in-variables; Total least squares

1. Introduction

A very common problem in statistics is to explain the effects of a covariate on a response (variable of interest). While the covariate is usually considered as a vector of scalars, nowadays, in many applications (for instance in climatology, remote sensing, linguistics, …) the data come from the observation of a continuous phenomenon over time or space: see Ramsay and Silverman (2002) or Ferraty and Vieu (2006) for examples. The increasing performances of measurement instruments permit henceforth to collect these data on dense grids and they cannot be considered anymore as variables taking values in $\mathbb{R}^p$. This necessitated to develop for this kind of data *ad hoc* techniques which have been popularized under the name of *functional data analysis* and have been deeply studied these last years (to get a theoretical and practical overview on functional data analysis, we refer to the books from Bosq, 2000; Ramsay and Silverman, 1997, 2002; Ferraty and Vieu, 2006).

Our study takes place in this framework of functional data analysis in the context of regression estimation evocated above. Thus, we consider here the case of a functional covariate while the response is scalar. To be more precise, we first consider observations $(X_i, Y_i)_{i=1,\dots,n}$, where the X_i's are real functions defined on an interval I of $\mathbb{R}$ with the assumption that it is square integrable over I. As usually assumed in the literature, we then work on the separable real Hilbert space $L^2(I)$ of such functions f defined on I such that $\int_I f(t)^2\,\mathrm{d}t$ is finite. This space is endowed with its usual inner product $\langle .,. \rangle$ defined by $\langle f, g\rangle = \int_I f(t)g(t)\,\mathrm{d}t$ for $f, g \in L^2(I)$, and the associated norm is noted $\|.\|_{L^2}$.

* Corresponding author. Tel.: +33 5 61556933; fax: +33 5 61556089.
E-mail addresses: cardot@enesad.inra.fr (H. Cardot), Christophe.Crambes@math.ups-tlse.fr (C. Crambes), akneip@uni-bonn.de (A. Kneip), Pascal.Sarda@math.ups-tlse.fr (P. Sarda).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2006.07.029

Now, the model we consider to summarize the link between covariates X_i and responses Y_i is a linear model introduced in Ramsay and Dalzell (1991) and defined by

$$Y_i = \int_I \alpha(t) X_i(t)\,\mathrm{d}t + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\alpha \in L^2(I)$ is an unknown functional parameter and ε_i, $i = 1, \ldots, n$ are i.i.d. real random variables satisfying $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}\left(\varepsilon_i^2\right) = \sigma_\varepsilon^2$. The functional parameter α has been estimated in various ways in the literature: see Ramsay and Silverman (1997), Marx and Eilers (1999) and Cardot et al. (1999,2003). Here, our final goal is to deal with the problem of estimating α in the case where $X_i(t)$ is corrupted by some unobservable error.

Before going further, let us note that there can be different ways to generate the curves X_i. One possibility is a fixed design, that is, $X_1, \ldots, X_n$ are fixed, non-random functions. Examples are experiments in chemical or engineering applications, where X_i corresponds to functional responses obtained under various, predetermined experimental conditions (see for instance Cuevas et al., 2002). In other applications one may assume a random design, where $X_1, \ldots, X_n$ are an i.i.d. sample. In any case, $Y_1, \ldots, Y_n$ are independent and the expectations always refer to the probability distribution induced by the random variables $\varepsilon_1, \ldots, \varepsilon_n$, only. In the case of random design, they thus formally have to be interpreted as conditional expectation given $X_1, \ldots, X_n$. This implies for instance that $\mathbb{E}(\varepsilon_i | X_i) = 0$ and $\mathbb{E}\left(\varepsilon_i^2 \middle| X_i\right) = \sigma_\varepsilon^2$.

In what precedes it is implicitly assumed that the curves X_i are observed without error (in model (1) all the errors are confined to the variable Y_i by the way of ε_i). Unfortunately, this assumption does not seem to be very realistic in practice, and many errors (instrument errors, human errors, ...) prevent to know $X_1, \ldots, X_n$ exactly. Furthermore, it is to be noted that in practice, the whole curves are not available, so we suppose in the following that the curves are observed in p discretization points $t_1 < \cdots < t_p$ belonging to I, that we will take equispaced. Taking from now on $I = [0, 1]$ in order to simplify the notations, we thus have $t_1 = 1/2p$, $t_j - t_{j-1} = 1/p$ for all $j = 2, \ldots, p$. Thus, we observe discrete noisy trajectories

$$W_i\left(t_j\right) = X_i\left(t_j\right) + \delta_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p, \tag{2}$$

where $\left(\delta_{ij}\right)_{i=1,\ldots,n,\ j=1,\ldots,p}$ is a sequence of independent real random variables, such that, for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$

$$\mathbb{E}\left(\delta_{ij}\right) = 0,$$

and

$$\mathbb{E}\left(\delta_{ij}^2\right) = \sigma_\delta^2.$$

The noise components δ_{ij} are not discrete realizations of continuous time "random noise" stochastic process and must be interpreted as random measurement errors at the finite discretization points (see e.g. Cardot, 2000; Chiou et al., 2003 for similar points of view).

The problem of the *errors-in-variables* linear model has already been studied in many ways in the case where the covariate takes values in $\mathbb{R}$ or $\mathbb{R}^p$, that is to say when it is univariate or multivariate. For instance, the maximum likelihood method has been applied to this context (see Fuller, 1987), and asymptotic results have been obtained (see for example Gleser, 1981). Because this problem is strongly linked to the problem of solving linear systems

$$\mathbf{A}\boldsymbol{x} \approx \mathbf{b},$$

where $\boldsymbol{x} \in \mathbb{R}^p$ is unknown, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{A}$ is a matrix of size $n \times p$, some numerical approaches have also been proposed. One of the most famous is the *total least squares* (*TLS*) method (see for example Golub and Van Loan, 1980; Van Huffel and Vandewalle, 1991).

Now, coming back to model (1), very few works have been done in the case of errors-in-variables: in a recent work Chiou et al. (2003) a two-step approach is proposed which consists in first smoothing the noisy trajectories in order to get denoised curves and then build functional estimators. The point of view adopted here is quite different and deals with the extension of the TLS approach in the context of the functional linear model.