PLS Regression with Functional Predictor and Missing Data

[1]Cristian Preda [2]Gilbert Saporta [3]M.H. Ben Hadj Mbarek

[1] Université des Sciences et Technologies de Lille 59650 Villeneuve d'Ascq, France.
E-mail: cristian.preda@polytech-lille.fr
[2]Chaire de Statistique appliquée, CEDRIC, CNAM, 292 Rue Saint Martin, 75141 Paris Cedex 03, France.
E-mail: gilbert.saporta@cnam.fr
[3]Institut Supérieur de Gestion de Sousse, Tunisie.
E-mail: benmbarekmhedi@yahoo.fr

Abstract: Time-average approximation and principal component analysis of the stochastic process underlying the functional data are the main ingredients for adapting NIPALS algorithm to estimate missing data in the functional context. The influence of the amount of missing data in the estimation of linear regression models is studied using the PLS method. A simulation study illustrates our methodology.
Keywords: functional data, missing data, PLS, functional regression models.

1 Introduction

Statistical methods for data representing functions or curves have received much attention in recent years. Such data, known in literature as functional data, have received in the last years a large interest for research, especially due to the difficulty to deal with infinite dimensional spaces in the context of classical multivariate methods. Examples of functional data can be found in several application domains such as medicine, economics, chemometrics and many others (Ramsay and Silverman, 2002) .

A well accepted model for functional data is to consider it as paths of a stochastic process $X = \{X_t, t \in [0, T]\}$ taking values into a Hilbert space of functions on some interval [0, T]. For example, a second order stochastic process $X=\{X_t, t \in [0, T]\}$ L2–continuous with sample paths in L_2 ([0, T]) can be used as model for describing the behavior of some quantitative parameter associated to a process observed on a time interval of length T.

Suppose that for each statistical unit ω we observe the associated curve X_{ω} (Figure 1 (a) provides an example) and a single real response Y_{ω}. We are interested in predicting Y_{ω} from X_{ω}. The linear functional regression model is the simplest approach to be considered and an important number of research papers in the functional data field are devoted to the estimation of the model,

$$Y=\int_0^T X_t\beta(t)dt + \varepsilon \qquad (1)$$

It is well known that the direct estimation of the regression coefficient function β using the least square criterion yields to an ill posed problem. Solutions based on elements derived from the principal component analysis of X have been proposed by Aguilera et al. (1997) and Cardot et al. (1999). These techniques are known in the literature as principal component regression (PCR). However, the choice of principal components is not an easy task, since one has to choose between robustness of the model (the most explanatories pc's) and his performances (the pc's the most correlated with the response). As an alternative to functional PCR, Preda and Saporta (2005) have extended Partial Least Squares (PLS) regression to the case of a functional predictor.

The aim of this paper is to provide a methodology for estimating linear regression models with functional predictor (X) in the presence of missing data. If missing data is quite a common concept in finite multivariate analysis (see Little and Rubin (1987)) that is not the case for functional data. In practice, a curve is generally observed in a finite number of time points, $0=t_0<t_1<\ldots< t_k = T$, and thus, with missing information. However, the true form of the curve can be approximated from the points $\{(t_i , X_{ti}), i=1, \ldots k\}$using interpolation or smoothing procedures (Aguilera et al. 1997).

We consider that a curve has missing data when one or several continuous part of the curve is missing, i.e observation was not possible. This situation occurs, for example, for instruments recording curves (spectrometers, oscilloscopes) that are out of service for some short time intervals. Figure 1 (b) provides an example of curve with missing data in two intervals of time.

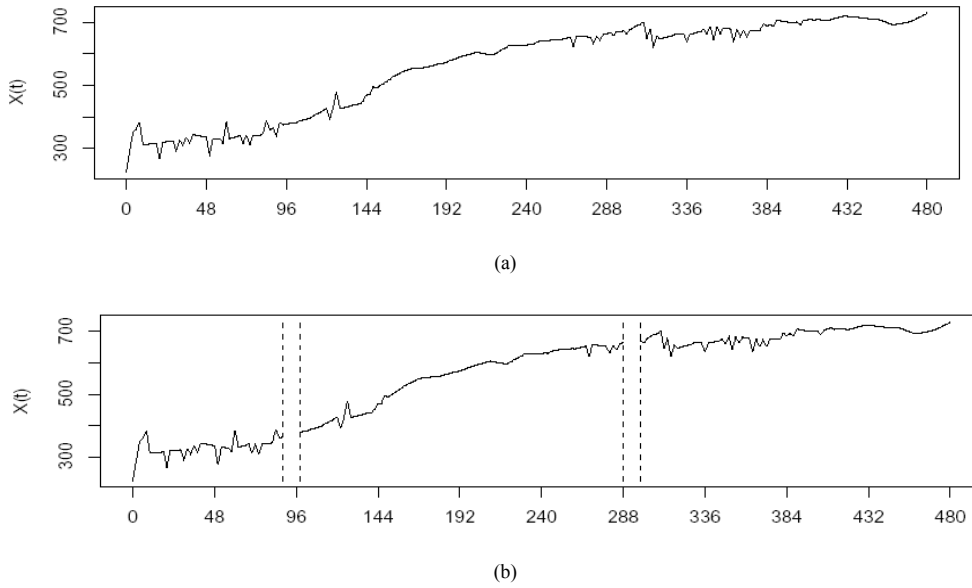

Figure 1 (a) Complete path. (b) Incomplete path

Our approach for dealing with missing data in the functional framework is to consider that the underlying process generating missing data is a jump stochastic process $M = \{M_t, t \in [0, T]\}$ with two states, {0, 1}, corresponding to the presence or absence of information.

For estimating the linear model (1) in presence of missing data we propose to use the PLS approach after time-average approximation (Preda (2000)) and imputation of missing data by the NIPALS algorithm (Tenenhaus, 1998).

The paper is organized as follows. In Section 2 we introduce the functional linear model for functional data and the PCR and PLS approaches. The process generating missing data as well as the methodology for applying the NIPALS algorithm for imputation is presented in Section 3. The Section 4 is devoted to a simulation study.

2 PLS Regression with Functional Data and Approximation

Let us consider the functional data as sample paths of a stochastic process $X=\{X_t, t\in [0, T]\}$ with continuous time and Y a random real variable defined on the same probability space as X. We assume that Y is centered and X is of second order, L_2 continuous and centered for each $t\in [0, T]$.

Under the least squares criterion, the estimation of the coefficient function β of linear regression model (1) is in general a distribution rather than a function of L_2 ([0, T]) (Saporta (1981)). This difficulty appears also in practice because one has generally more predictors than the number of observations. Regression on principal components

(PCR) of X (Aguilera et al., 1997) and PLS approach (Preda and Saporta (2005)) provide efficient solutions to this problem. We will consider here only the PLS approach.

2.1 Partial least squares regression for functional data

The basic idea of PLS approach is to construct a set of uncorrelated random variables $\{T_i, i \geqslant 1\}$ (PLS components) in the linear space spanned by X, taking into account the correlation between Y and X. Replacing the least squares criterion with that of maximal covariance between X and Y,

$$\max_{w} \ \mathrm{cov}^2(Y, \int_0^T X_t\, w(t)dt)$$

The PLS regression offers a good alternative to PCR (Preda and Saporta , 2005). The first PLS component is given by $T_1 = \int_0^T X_t\, w(t)dt$ and further PLS components are obtained by maximizing the covariance criterion between the residuals of both Y and X_t with the previous components.

The PLS approximation is given by

$$\hat{Y}_{PLS(k)} = \sum_{i=1}^{k} c_i T_i \ = \int_0^T X_t\, \beta_{PLS(k)}(t)dt \tag{2}$$

As in the finite multivariate setting (de Jong, 1993), in the functional context PLS fits closer than PCR, i.e. $R^2(Y, \hat{Y}_{PCR(k)}) \leqslant R^2(Y, \hat{Y}_{PLS(k)})$.

2.2 Time average approximation

The principal components analysis of X is often realized by approximating the principal factors in a finite dimensional space of functions. One of the approximations, which is convenient in presence of missing data, is the time-average approximation developed in Preda (2000). This approximation, easy to put in practice, consists into approximate X by a stochastic process with whose the sample paths are constant piecewise functions. If $\Delta=\{0= t_0 < t_1 <\ldots < t_p=T\}$ is a discretisation of [0, T] then the time-average approximation of X is given by X^{Δ} defined by

$$X_t^{\Delta} = m_i = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} X_t\, dt, \quad \forall t \in [t_{i-1}, t_i\,], \ i=1, \ldots, p \tag{3}$$

Properties of this approximation with respect to the accuracy of the approximations provided by the elements derived from principal components analysis are presented in Preda (2000). Let observe that the principal component analysis in this case is equivalent with the principal component analysis of the set of variables $\{m_i, i=1, \ldots, p\}$ using as metric diag $(t_1-t_0, \ldots., t_p-t_{p-1})$. The principal factors f_i of the process X are approximated by constant piecewise functions f^{Δ}_i obtained from the principal factors of the set $\{m_i, i=1, \ldots, p\}$ and so are for the principal components, $\xi^{\Delta}_{i.}$

The functional PCR regression of Y on X is then approximated by the PCR of Y on the set $\{\xi^{\Delta}_{i.}, i=1, \ldots, k, k\leqslant p\}$.

In the same way, the PLS regression of Y on X is approximated by the PLS regression of Y on the set of variables $\{\sqrt{t_i - t_{i-1}} \times m_i\ , 1, \ldots, p\}$ (Preda and Saporta, 2005).

3 Missing Data for Functional Data and NIPALS Algorithm

At our knowledge, there are not works on dealing with missing data for functional variables. We can observe that when the situation occurs, it is often question of the end of the curve and thus imputation of missing data is

synonym of time series prediction.

3.1 Missing data model

In our approach we consider that the missing information could occur in any continuous time interval of [0, T]. Thus, a curve can miss information of a set of intervals $[a_1, b_1]$, …, $[a_m, b_m]$. Of course, the number of these intervals is random as well as their length. One possible model for missing data in this context is to consider an underlying jump continuous time process M_t with two states, 0 and 1, with the following signification

$$M_t = \begin{cases} 0, & \text{if } X_t \text{ is observed at time t,} \\ 1, & \text{otherwise.} \end{cases} \quad (4)$$

Thus, to each curve ω of X corresponds a curve ω_M of M. A curve ω_M that corresponds to the "0" constant function means that the curve ω is completely observed.

In the multivariate finite case, it is usually to speak about the ratio of the missing data in the whole dataset. In the functional context, we can extend this notion to the ratio of the sum of the length of the intervals $[a_i, b_i]$ within [0, T] for all available curves. However, if this ratio has some interpretation when the missing data is "completely at random" (see Little and Rubin, 1987), it is difficult to justify this measure in the case of functional data.

Inspired by the reliability theory of repairable systems, we propose as measure for quantify the missing information the mean time of missing observation (MTMO) defined by

$$\text{MTMO} = \frac{1}{T}\int_0^T U(t)dt, \quad (5)$$

where U (t) is the probability that the process X is not observable at the instant t, i.e. U (t) = P (M_t = 1). Obviously, the simplest model for M is a two state markovian process with exponential times for each state. Considering that M_0=0 and the rate parameters describing the system are λ (for state 0) and μ (for state 1) then one can show (Iosifescu et al (2007) that

$$\text{MTMO} = \frac{\lambda}{\lambda+\mu} - \frac{\lambda}{(\lambda+\mu)^2 T} \times \left(1 - e^{-\frac{\lambda+\mu}{T}}\right) \quad (6)$$

For example, for T=1, λ=1 and μ=100 we have MTBO = 0.009802 that means that the process is unobservable about of 1% of time.

3.2 Estimation of missing data by the NIPALS algorithm

For each curve ω let consider the set of intervals $[a_1(\omega), b_1(\omega)]$, …$[a_m(\omega), b_m(\omega)]$ corresponding to missing data (eventually empty). Let $\Delta=\{0=t_0 < t_1 < \ldots < t_p=T\}$ be an equidistant discretization of [0, T] with $\delta = t_{i+1} - t_i$ such that each length of data missing intervals is multiple of δ. Then, the time average approximation can be used by the NIPAL algorithm in order to predict the missing values of the variables defined by (3), $\{m_i, i=1, \ldots, p\}$.

Based on the estimation of simple linear regression models with missing data, the NIPALS algorithm provides estimation for the principal components and principal factors of the principal component analysis of X. Therefore, by the reconstruction formula of principal component analysis, one obtains values for the missing data and PLS regression model can be estimated.

4 Simulation Study

Let consider Y be a real random variables defined by the linear regression model

$$Y = \int_0^1 X_t \beta(t)dt + \varepsilon , \tag{7}$$

where X is the standard Brownian motion on [0, 1], $\beta(t) = 3t^3$, $t \in [0, 1]$ and ε is the error term such that V (ε) =0, 1. Notice that V (Y) =0, 5 and thus R^2 =0.8.

We generate n=100 curves representing the sample paths of X observed on a discretization Δ_1 of the interval [0, 1] in 1000 equidistant intervals. The Simpson quadrature method provides the values of Y for each one of curves.

In order to smooth the local variation of curves, a time-average approximation is performed on data using a discretization Δ_2 with 100 equidistant intervals. Then, the PLS estimation, $\beta_{PLS,}$ of the regression coefficient function β is obtained with three PLS components and is represented in Figure 2 (a) .

Missing data is now simulated for several values of MTMO. We use for simulation a two state markovian jump process with exponential times for the two states as in (6). The exponential distribution is simulated with a precision of 1/1000, thus the change points belongs to Δ_1. In Table 1 are presented the performances (measured by R^2) of the PLS regression after time-average approximation and estimation of the missing data using the NIPALS algorithm for several values of MTMO (λ and μ).

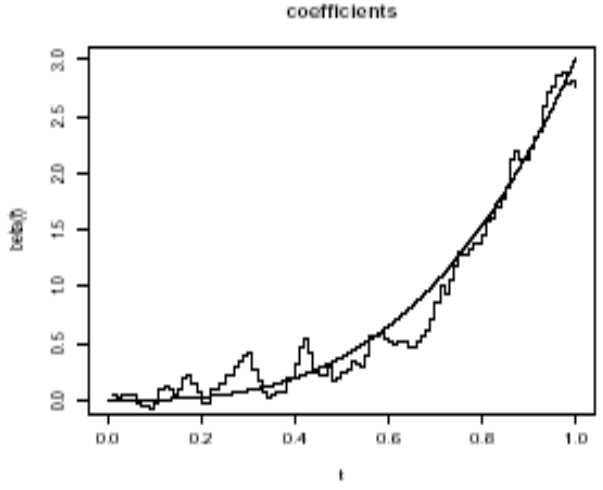

Figure 2 Regression coefficient function with complete data

Table 1 Performances of the PLS regression with missing data

Parameters	MTMO	R^2
Complete data	0	0.7645
λ=1, μ=100	0.00980	0.7263
λ=1, μ=50	0.01922	0.7288
λ=1, μ=20	0.04535	0.7144
λ=2, μ=20	0.08677	0.6625
λ=2, μ=10	0.15277	0.6218
λ=2, μ=5	0.24493	0.4872

References

[1] Aguilera A.M., Ocana F., Valderrama M.J. (1997) *An approximated principal component prediction model for continous-time stochastic process*, Applied Stochastic Models and Data Analysis, Vol. 13, 61-72.

[2] Cardot H., Ferraty F., Sarda P. (1999) Functional linear model, *Statist. Prob. Lett.,* 45, 11-22.

[3] de Jong S. (1993) *PLS fits closer than PCR*, Chemometrics, Vol. 7, 551-557.

[4] Iosifescu M., Limnios N., Oprisan G. (2007) Modèles Stochastiques, Hermes Science.

[5] Little R.J.A., Rubin D.B. (1987) *Statistical analysis with missing data*, Wiley, New York 1987.

[6] Preda C. (2000) *Approximation par moyennage de l'analyse en composantes principales d'un processus stochastique*, Comptes rendus de l'Académie des Sciences de Paris, T. 330, Série I, 1-6.

[7] Preda C. Saporta G. (2005) PLS regression on a stochastic process, *Computational Statistics and Data Analysis*, 48, 149-158.

[8] Ramsay J.O., Silverman B.W. (2002) *Applied Functional Data Analysis : Methods and Case Studies*, Springer Series in Statistics, Springer-Verlag, New York.

[9] 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.

[10] Tenenhaus M. (1998*) La régression PLS Théorie et pratique*. Editions Technip.