Journal of Statistical Planning and Inference 162 (2015) 75–87

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Nonparametric regression estimation for functional stationary ergodic data with missing at random

Nengxiang Ling[a,*], Longlong Liang[a], Philippe Vieu[b]

[a] School of Mathematics, Hefei University of Technology, Hefei, 230009, PR China
[b] Institut de Mathématiques, Université Paul Sabatier, Toulouse, France

ARTICLE INFO

Article history:
Received 18 June 2014
Received in revised form 15 November 2014
Accepted 2 February 2015
Available online 16 February 2015

Keywords:
Missing at random
Functional data analysis
Convergence in probability
Asymptotic normality
Ergodic processes
Regression operator

ABSTRACT

In this paper, we investigate the asymptotic properties of the estimator for the regression function operator whenever the functional stationary ergodic data with missing at random (MAR) are considered. Concretely, we construct the kernel type estimator of the regression operator for functional stationary ergodic data with the responses MAR, and some asymptotic properties such as the convergence rate in probability as well as the asymptotic normality of the estimator are obtained under some mild conditions respectively. As an application, the asymptotic $(1-\zeta)$ confidence interval of the regression operator is also presented for $0 < \zeta < 1$. Finally, a simulation study is carried out to compare the finite sample performance based on mean square error between the classical functional regression in complete case and the functional regression with MAR.

1. Introduction

It is well known that a common problem in statistics is to explore the relationship between a variable of interest Y (called response variable) and an explanatory variable X. This paper focuses on the case where Y is a scalar response variable and X is of functional nature which takes values in some abstract infinite dimensional space $(\mathscr{H}, d(.,.))$, where $d(.,.)$ is an associated semi-metric. In this framework, let us consider the following functional nonparametric regression model:

$$Y = r(X) + \varepsilon, \tag{1.1}$$

where $r(.)$ is an unknown smooth functional regression operator from $\mathscr{H}$ to $\mathbb{R}$, and ε is the random error with $E\varepsilon = 0$ and $0 < Var(\varepsilon) < \infty$.

Compared with the classical nonparametric regression framework that the explanatory variable is a real or finite dimensional case, model (1.1), where the explanatory variables X are often curves or surfaces, is widely applied in many fields such as in medicine, economics, environmetrics, chemometrics and others, The reason is that the data we observed or collected in these fields are exceptionally high-dimensional or even functional. The statistical problems involving the modelization of functional random variables have received an increasing interest in recent literature, and we can see Ramsay and Silverman (2005) for parametric models and Ferraty and Vieu (2000) for nonparametric regression model, respectively. In fact, model (1.1) was first introduced by Ferraty and Vieu (2000). Then, the theory and methods in this research field are well developed, see, for instance, Ferraty and Vieu (2002, 2003, 2004) and Ferraty et al. (2006), as attested by the monograph

* Corresponding author.
E-mail addresses: hfut.lnx@163.com (N. Ling), lianglong65@163.com (L. Liang), philippe.vieu@math.univ-toulouse.fr (P. Vieu).

http://dx.doi.org/10.1016/j.jspi.2015.02.001
0378-3758/

by Ferraty and Vieu (2006) and the references therein. All the results involved above are in the case that the samples are observed completely. However, in many practical works such as sampling survey, pharmaceutical tracing test and reliability test and so on, some pairs of observations may be incomplete, which is often called the case of missing data. Many examples of missing data and its statistical inferences for regression model can be found in statistical literature when explanatory variables are of finite dimensionality. For example, we can quote: Cheng (1994), Little and Rubin (2002), Nittner (2003), Tsiatis (2006), Liang et al. (2007), Efromovich (2011a,b) and references therein for details. When explanatory variables are in the case of infinite dimensionality or it is of functional nature, only very few literature was reported to investigate the statistical properties of functional nonparametric regression model for missing data. Recently, Ferraty et al. (2013) first proposed to estimate the mean of a scalar response based on an i.i.d. functional sample in which explanatory variables are observed for every subject, while the response variables are missing at random by happenstance for some of them. It generalized the results in Cheng (1994) to the case where the explanatory variables are of functional nature.

Inspired by all the papers above, our work in this paper aims to contribute to the research on functional nonparametric regression model (1.1), by studying the estimation of regression operator based on missing data. More precisely, we construct the estimator of the regression operator of a scalar response and the functional covariate which are assumed to be sampled from a stationary and ergodic process. Meanwhile, the response variables are MAR but not the covariates are missing. Then, the asymptotic properties of the estimator are obtained under some mild conditions. Besides the infinite dimensional character of the data with MAR, we avoid here the widely used strong mixing conditions, which are still to be verified or even fail to hold for some process induced, see the discussions in Laib and Louani (2010, 2011) for details. As far as we know, the estimation of the nonparametric regression operator combining missing data and stationary ergodic processes with functional nature has not been studied in the statistical literature.

The organization of the paper is as follows. In Section 2, we describe some assumptions on the model (1.1) and construct precisely the estimator of $r(.)$ based on the functional stationary ergodic data with MAR, while in Section 3 we present the main results of our work. In Section 4, we illustrate our methodology by a simulation study to compare the classical nonparametric functional model in complete and the model with MAR. Finally, the proofs of the main results are postponed to Section 5.

2. Model assumptions and estimators

2.1. Functional MAR modeling and estimates

In this section, we focus on estimation of regression operator $r(.)$. Let $\{(X_i, Y_i), 1 \le i \le n\}$ be a sequence of stationary and ergodic random samples from the model (1.1). Thus, we have that

$$Y_i = r(X_i) + \varepsilon_i, \quad (i = 1, 2, \dots, n), \tag{2.1}$$

where $E\varepsilon_i = 0$ and $0 < Var(\varepsilon_i) < \infty$. It is clear that in the case of complete data, a well-known N-W kernel-type estimator of $r(x)$ is given by

$$\widetilde{r}_n(\chi) = \frac{\sum_{i=1}^{n} Y_i K_h(d(\chi, X_i))}{\sum_{i=1}^{n} K_h(d(\chi, X_i))}, \quad \chi \in \mathscr{H}, \tag{2.2}$$

where $K_h(u) = K(u/h)$, $K(.)$ is a real-valued kernel function from $[0, \infty)$ into $[0, \infty)$, $h = h_n > 0$ is a smoothing parameter satisfying $h_n \to 0$ as $n \to \infty$ and $d(., .)$ is a semi-metric on $\mathscr{H}$. However, in missing mechanism with MAR for the response variable, an available incomplete sample of size n from (X, Y, δ) is $\{(X_i, Y_i, \delta_i), 1 \le i \le n\}$, where X_i is observed completely, $\delta_i = 1$ if Y_i is observed, and $\delta_i = 0$ otherwise. Meanwhile the Bernoulli random variable δ is satisfied with

$$P(\delta = 1|X = \chi, Y = y) = P(\delta = 1|X = \chi) = p(\chi),$$

where $p(\chi)$ is a function operator, which is called the conditional probability of the observing response given the predictor and is often unknown. This mechanism shows that δ and Y are conditionally independent given X. Therefore, similar to Cheng (1994) in finite dimensionality case, we present the estimator of $r(\chi)$, which is given by

$$\widehat{r}_n(\chi) = \frac{\sum_{i=1}^{n} \delta_i Y_i K\left(\frac{d(\chi, X_i)}{h}\right)}{\sum_{i=1}^{n} \delta_i K\left(\frac{d(\chi, X_i)}{h}\right)} = \frac{\widehat{r}_{n,2}(\chi)}{\widehat{r}_{n,1}(\chi)}, \tag{2.3}$$

where $K(.)$, $h = h_n$ and $d(., .)$ are defined as that in (2.2) respectively, and

$$\widehat{r}_{n,j}(\chi) = \frac{1}{nE(\Delta_1(\chi))} \sum_{i=1}^{n} Y_i^{j-1} \delta_i \Delta_i(\chi), \quad j = 1, 2 \tag{2.4}$$