Journal of Multivariate Analysis 147 (2016) 168–182

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Efficiency in multivariate functional nonparametric models with autoregressive errors

S. Dabo-Niang [a,c], S. Guillas [b], C. Ternynck [a,*]

[a] Laboratoire LEM, Université de Lille, Villeneuve-d'Ascq, France
[b] University College London, London, United Kingdom
[c] Modal Team INRIA, Lille, France

ARTICLE INFO

Article history:
Received 15 January 2015
Available online 4 February 2016

AMS 2000 subject classifications:
62G08
62M10

Keywords:
Autoregressive process
Functional data
Kernel regression
Pre-whitening
Time series

ABSTRACT

In this paper, we introduce a new procedure for the estimation in the nonlinear functional regression model where the explanatory variable takes values in an abstract function space and the residual process is autocorrelated. Moreover, we consider the case where the response variable takes its values in $\mathbb{R}^d$. The procedure consists in a pre-whitening transformation of the dependent variable based on the estimated autocorrelation. We establish both consistency and asymptotic normality of the regression function estimate. For kernel methods encountered in the literature, the correlation structure is commonly ignored (the so-called "working independence estimator"); we show here that there is a strong benefit in taking into account the autocorrelation in the error process. We also find that the improvement in efficiency can be large in our functional setting, up to 25% in the presence of high autocorrelation levels. We observe that the additional step of iterating the fitting process actually deteriorates the estimation. We illustrate the skills of the methods on simulations as well as on application on ozone levels over the US.

1. Introduction

The use of functional data analysis methods is spreading in statistics due to the availability of high frequency data and of new mathematical strategies to deal with such statistical objects. The field is known as Functional Data Analysis (FDA). Applications of FDA are growing across fields as diverse as energy studies [2], linguistics [3], atmospheric chemistry [29], and human vision [28]. Functional variables are often curves, but surfaces and manifolds have also been considered recently (e.g., [23,34]). For an introduction to this field, along with illustrations and applications, see [31]. For a mathematical treatment of nonparametric methods suitable for functional regression, refer to Ferraty and Vieu [21], and for a survey of the state of the art in FDA theory, see [8].

Kernel-based methods are often used to estimate the regression operator. This approach yields almost sure consistency in the case of an independent sample [19] or an α-mixing sample [16,17], but also asymptotic normality in the independent case [18] with exact computation of all the constants for its precise use in practice. Masry [27] established the asymptotic normality of the nonparametric regression estimator for strongly mixing processes albeit with abstract expressions of the constants so this is more challenging to use in practice. Delsol [12,13] generalized the results of Ferraty et al. [18] to the case of an α-mixing dataset.

* Correspondence to: Universite de Lille, Faculté de Médecine - Pôle Recherche (2ème étage), CERIM, 1 place de Verdun, 59045 Lille Cedex, France.
E-mail addresses: sophie.dabo@univ-lille3.fr (S. Dabo-Niang), s.guillas@ucl.ac.uk (S. Guillas), ternynck.camille@gmail.com, camille.ternynck@univ-lille2.fr (C. Ternynck).

http://dx.doi.org/10.1016/j.jmva.2016.01.007
0047-259X/

In this paper, we consider the regression of a multivariate random variable onto a functional random variable. The estimation of the regression function is tackled by means of a nonparametric kernel approach. The existing kernel regression estimators dealing with functional explanatory variables are for scalar response; we have not found existing research on functional nonparametric modeling for multivariate response. With multivariate explanatory variables and a multivariate response, Xiang et al. [36] proposed a kernel estimate of the regression function. Our regression model below is an extension of Xiang et al. [36]:

$$\mathbf{Y}_t = \mathbf{m}(X_t) + \mathbf{u}_t, \quad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{Y}_t = (Y_{t,1}, \ldots, Y_{t,d})^\top \in \mathbb{R}^d$, $\mathbf{m}(X_t) = (m_1(X_t), \ldots, m_d(X_t))^\top$, the explanatory variable is functional (that is, X_t takes values in some possibly infinite-dimensional space), $\mathbf{u}_t = (u_{t,1}, \ldots, u_{t,d})^\top$. Moreover, the stationary residual process $\mathbf{u}_t$ is autocorrelated and independent of X_t. We do not necessarily assume that $(X_t, \mathbf{Y}_t)_t$ is strictly stationary, second order stationarity suffices.

Although, for the kernel methods proposed in the literature, it is generally better to ignore the correlation structure entirely (the so-called "working independence estimator", e.g. [33,26]), we show here that taking into account the autocorrelation of the error process helps improve the estimation of the regression function.

We extend the kernel-based procedure proposed by Xiao et al. [37] for estimating $m(x)$ in the time series regression model for multivariate explanatory variables x to a functional setting. Xiao et al. [37] showed that their procedure is more efficient than the conventional local polynomial method. The main idea is to transform the original regression model, so that this transformed regression has a residual term that is uncorrelated. This transformation depends on the function $\mathbf{m}(\cdot)$ and on the parameters of the autoregressive representation of $\mathbf{u}$, since the regression function is nonlinear. The error correlation structure is assumed to have an autoregressive representation. Firstly, the parameters of the autoregressive representation are estimated. In a second step, a transformation $\widehat{\underline{\mathbf{Y}}}_t$ of the dependent variable $\mathbf{Y}_t$ is constructed by plugging in the estimated autocorrelation parameter. Finally, the estimation of $\mathbf{m}$ is carried out on this filtered series $\widehat{\underline{\mathbf{Y}}}_t$.

The remainder of the paper is organized as follows. In Section 2, we introduce the estimation method as well as the assumptions. We then provide asymptotic results for the estimator proposed. Section 3 is devoted to a simulation case study and an illustration of our method for ozone levels over the US. The conclusion is done in Section 4 while the proofs are given in the Appendix A.

2. Assumptions and main results

Suppose that we have a sample $\{(X_1, \mathbf{Y}_1), \ldots, (X_T, \mathbf{Y}_T)\}$, where for each $t \in \{1, \ldots, T\}$, X_t is a random variable taking its values in a semi-metric space $(\mathcal{C}, d)$ of infinite dimension and $\mathbf{Y}_t \in \mathbb{R}^d$ is the response from the nonparametric regression (1). We assume that the residual process $\mathbf{u}_t \in \mathbb{R}^d$ is stationary, has mean $\mathbf{0}$ with cross-covariance (auto-covariance in the univariate case) $\gamma_{\mathbf{u}}$ and has the following invertible linear process representation (with bounded coefficients):

$$\mathbf{u}_t = \sum_{k=0}^{\infty} \Psi_k \mathbf{e}_{t-k} = \Psi(L)\mathbf{e}_t, \tag{2}$$

where $\Psi_0 = I$ is the identity matrix, $\Psi(L) = \sum_{k=0}^{\infty} \Psi_k L^k$ is a $d \times d$ matrix in the lag operator L ($L^k(\mathbf{e}_t) = \mathbf{e}_{t-k}$), the (i,j)th element of $\Psi(L)$ is $\psi_{ij}(L) = \sum_{k=0}^{\infty} c_{ij}(k)L^k$, the random vectors $\mathbf{e}_t \in \mathbb{R}^d$ form a white noise process with mean $\mathrm{E}(\mathbf{e}_t) = \mathbf{0}$, $\mathrm{E}(\mathbf{e}_t\mathbf{e}_t^\top) = \Sigma_{\mathbf{e}}$ is a positive definite matrix, $\mathrm{E}(\mathbf{e}_t\mathbf{e}_{t+k}) = \mathbf{0}$ for $k \neq 0$ and $\mathrm{E}\left(|e_{t,j}|\right) < \infty$ for all $j \in \{1, \ldots, d\}$.

Let $\Psi(L)^{-1} = \Pi(L) = I - \sum_{k=1}^{\infty} \Pi_k L^k$ with $\Pi_0 = I$, or as done for Ψ let the (i,j)th element of $\Pi(L)$ be $\pi_{ij}(L) = \sum_{k=0}^{\infty} a_{ij}(k)L^k$. We then have the infinite autoregressive representation

$$\Pi(L)\mathbf{u}_t = \mathbf{e}_t. \tag{3}$$

Note that stationary, causal and invertible vector ARMA processes

$$\mathbf{u}_t - \sum_{k=1}^{p} \Phi_k \mathbf{u}_{t-k} = \mathbf{e}_{t-k} - \sum_{k=1}^{q} \Theta_k \mathbf{e}_{t-k}$$

can be represented as in (2)–(3) if all roots of $\det\{\Phi(L)\}$ and $\det\{\Theta(L)\}$ are greater than 1 in absolute value.

Here, we consider a truncated version of $\Pi(L)$ at order Q, i.e., $\Pi(L) = I - \sum_{k=1}^{Q} \Pi_k L^k$, where the truncation parameter Q is large enough. Applying $\Pi(L)$ to the regression in Eq. (1), we obtain $\Pi(L)\mathbf{Y}_t = \Pi(L)\mathbf{m}(X_t) + \mathbf{e}_t$. Then let the regression model $\underline{\mathbf{Y}}_t = \mathbf{m}(X_t) + \mathbf{e}_t$, with $\underline{\mathbf{Y}}_t = \mathbf{Y}_t - \sum_{k=1}^{Q} \Pi_k L^k\{\mathbf{Y}_t - \mathbf{m}(X_t)\}$, so the error term in this transformed model is now uncorrelated. The matrix of coefficients $\{\Psi_k\}_{k=0}^{\infty}$ and the regression function $\mathbf{m}(\cdot)$ are unknown, except for the fact that $\mathbf{m}(\cdot)$ is a smooth function. If $\underline{\mathbf{Y}}_t$ were known then a nonparametric kernel regression of $\underline{\mathbf{Y}}_t$ on X_t would be more efficient than the conventional kernel estimation. In this work, we employ a Nadaraya–Watson estimator as introduced in [20,27,11] where