Journal of Multivariate Analysis 146 (2016) 301–312

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Multivariate functional linear regression and prediction

Jeng-Min Chiou *, Ya-Fang Yang, Yu-Ting Chen

Institute of Statistical Science, Academia Sinica, Nankang, Taipei 11529, Taiwan

ARTICLE INFO

Article history:
Received 1 January 2015
Available online 23 October 2015

AMS 2010 subject classifications:
62H25
62M09

Keywords:
Functional prediction
Functional principal component analysis
Functional regression
Multivariate functional data
Stochastic processes

ABSTRACT

We propose a multivariate functional linear regression (mFLR) approach to analysis and prediction of multivariate functional data in cases in which both the response and predictor variables contain multivariate random functions. The mFLR model, coupled with the multivariate functional principal component analysis approach, takes the advantage of cross-correlation between component functions within the multivariate response and predictor variables, respectively. The estimate of the matrix of bivariate regression functions is consistent in the sense of the multi-dimensional Gram–Schmidt norm and is asymptotically normally distributed. The prediction intervals of the multivariate random trajectories are available for predictive inference. We show the finite sample performance of mFLR by a simulation study and illustrate the method through predicting multivariate traffic flow trajectories for up-to-date and partially observed traffic streams.

1. Introduction

Functional regression analysis is widely used to describe the relationship between response and predictor variables when at least one of the variables contains a random function (see Cuevas [11], Ferraty and Vieu [17], Horváth and Kokoszka [23], Müller [29], and Ramsay and Silverman [34] for excellent overviews). There has been intensive literature in the functional linear regression (FLR) models of the type of a scalar response and a functional predictor, the simple FLR. Two methods are typically used to address the simple FLR model. The most popular one is based on functional principal component analysis (FPCA) (e.g., Cardot et al. [5] and Hall and Horowitz [21]). The other is based on penalized regularization such as the penalized B-splines (Li and Hsing [26]) and the reproducing kernel Hilbert space approaches (Yuan and Cai [37]).

This simple FLR model was extended to nonlinear and semiparametric functional regression models to accommodate the case of multiple functional predictors or the situation when the relationship between response and predictor variables is not linear. These include the generalized single and multiple index FLR models with known or unknown link (Amato et al. [1], Chen et al. [6], Ferraty et al. [15], James [24], Müller and Stadtmüller [30]), the additive functional regression model (Febrero-Bande, Gonzalez-Manteiga [14], Ferraty and Vieu [18], Goia and Vieu [20]), the semi-functional partial linear regression (Aneiros-Perez and Vieu [3]), FLR with derivatives as supplementary covariates (Goia [19]), and time series prediction (Mas and Pumo [27]).

FLR models with a single random function as the response were introduced by Ramsay and Dalzell [33]. These include functional response models with scalar predictors (Chiou et al. [9,10], Faraway [13]) and the functional linear regression model with a functional response and a functional predictor (Ferraty et al. [16], Yao et al. [36]). Generalizations of functional response models include functional additive models (Müller and Yao [31]) and the functional mixture prediction model (Chiou [7]). Functional response models with more than one functional predictors were discussed in Matsui et al. [28]. More

* Corresponding author.
E-mail address: jmchiou@stat.sinica.edu.tw (J.-M. Chiou).

http://dx.doi.org/10.1016/j.jmva.2015.10.003
0047-259X/

recently, the functional response additive model estimation (Fan et al. [12]) and the functional errors-in-variable (Radchenko et al. [32]) approaches consider univariate functional responses and multivariate functional predictor variables, which are interesting methods for functional response models. Albeit the extensive development of functional regression models, functional response models with multivariate random functions as the response variable have not been discussed in the literature.

We develop a multivariate FLR (mFLR) model in which both the response and predictor variables contain multivariate random trajectories and are contaminated with measurement errors. The mFLR model takes the advantage of component dependency of multivariate random functions and accommodates incomparable magnitudes of variation among the component functions of the response and predictor variables, respectively. We discuss the existence and uniqueness of the estimates of bivariate regression functions, obtain the asymptotic properties of the estimators, and construct relevant pointwise and simultaneous prediction intervals for the predictive inference. We illustrate the finite sample performance of the proposed mFLR approach through a simulation study and apply the method to predict future multivariate traffic flow trajectories for an up-to-date and partially observed traffic streams in intelligent transportation systems.

The remainder of this article is organized as follows. In Section 2, we present the proposed multivariate FLR (mFLR) model. In Section 3, we discuss estimation of the regression model and prediction of future multivariate trajectories with prediction intervals. In Section 4, we derive the asymptotic properties of the mFLR model. In Section 5, we present the numerical results of a simulation study and a real-life application to multivariate traffic-flow data. Technical details and information on the estimation process are compiled in Appendices A.1–A.5. More technical details and numerical results are provided in the online Supplement (see Appendix B).

2. Multivariate functional linear regression model

2.1. Preliminaries

Let $\{X_l\}_{1\le l\le p}$ and $\{Y_k\}_{1\le k\le d}$ be the sets of random functions, corresponding to the predictor and response variables, with each X_l in $L_2(\mathcal{S})$ and Y_k in $L_2(\mathcal{T})$, where $L_2(\cdot)$ is a Hilbert space of square-integrable functions with respect to Lebesgue measures ds and dt on closed intervals $\mathcal{S}$ and $\mathcal{T}$.

Further, let $\boldsymbol{X}(s) = \left(X_1(s), \dots, X_p(s)\right)^\top$ be a vector in a Hilbert space of p-dimensional vectors of functions in $L_2(\mathcal{S})$, denoted by $\mathbb{H}_1 = L_2^p(\mathcal{S})$. Assume $\boldsymbol{X}(s)$ has a smooth mean function $\boldsymbol{\mu}^X(s) = (\mu_1^X(s), \dots, \mu_p^X(s))^\top$, $\mu_l^X(s) = \mathrm{E}X_l(s)$, and covariance function $\boldsymbol{G}^X(s_1,s_2) = \left\{G_{jl}^X(s_1,s_2)\right\}_{1\le j,l\le p}$, $G_{jl}^X(s_1,s_2) = \mathrm{cov}\left(X_j(s_1), X_l(s_2)\right)$. Similarly, $\boldsymbol{Y}(t) = (Y_1(t), \dots, Y_d(t))^\top$ in $\mathbb{H}_2 = L_2^d(\mathcal{T})$ has a smooth mean function $\boldsymbol{\mu}^Y(t) = (\mu_1^Y(t), \dots, \mu_d^Y(t))^\top$, $\mu_k^Y(t) = \mathrm{E}Y_k(t)$, and covariance function $\boldsymbol{G}^Y(t_1,t_2) = \left\{G_{km}^Y(t_1,t_2)\right\}_{1\le k,m\le d}$, $G_{km}^Y(t_1,t_2) = \mathrm{cov}\left(Y_k(t_1), Y_m(t_2)\right)$. In addition, let the diagonal matrices be $\boldsymbol{D}^X(s) = \mathrm{diag}\left(v_1^X(s)^{1/2}, \dots, v_p^X(s)^{1/2}\right)$, where $v_l^X(s) = G_{ll}^X(s,s)$ and $\boldsymbol{D}^Y(s) = \mathrm{diag}\left(v_1^Y(t)^{1/2}, \dots, v_d^Y(t)^{1/2}\right)$, where $v_k^Y(t) = G_{kk}^Y(t,t)$. The inner product of any functions f and g in $L_2(\mathcal{S})$ is $\langle f, g\rangle = \int_{\mathcal{S}} f(s)g(s)\mathrm{d}s$, with the norm $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$. The inner product of any functions $\boldsymbol{f} = (f_1, f_2, \dots, f_p)^\top$ and $\boldsymbol{g} = (g_1, g_2, \dots, g_p)^\top$ in $\mathbb{H}_1$ is $\langle \boldsymbol{f}, \boldsymbol{g}\rangle_{\mathbb{H}_1} = \sum_{l=1}^p \langle f_l, g_l\rangle$, and the norm $\|\cdot\|_{\mathbb{H}_1} = \langle\cdot,\cdot\rangle_{\mathbb{H}_1}^{1/2}$. The inner product of two functions in $\mathbb{H}_2$, $\langle\cdot,\cdot\rangle_{\mathbb{H}_2}$, is defined in the same way.

To accommodate incomparable magnitudes of variation between the component functions $\{X_l(s)\}$ (resp. $\{Y_k(t)\}$), we take the transformation approach of Chiou et al. [8]. Let $\boldsymbol{X}^Z(s) = (X_1^Z(s), \dots, X_p^Z(s))^\top = \boldsymbol{D}^X(s)^{-1}\{\boldsymbol{X}(s) - \boldsymbol{\mu}^X(s)\}$ (resp. $\boldsymbol{Y}^Z(t) = (Y_1^Z(t), \dots, Y_d^Z(t))^\top = \boldsymbol{D}^Y(t)^{-1}\{\boldsymbol{Y}(t) - \boldsymbol{\mu}^Y(t)\}$). It follows that $\boldsymbol{X}^Z(s)$ (resp. $\boldsymbol{Y}^Z(t)$) has a mean of 0 and covariance function $\boldsymbol{C}^X(s_1,s_2) = \{C_{jl}^X(s_1,s_2)\}$ (resp. $\boldsymbol{C}^Y(t_1,t_2) = \{C_{km}^Y(t_1,t_2)\}$), where $C_{jl}^X(s_1,s_2) = \mathrm{E}\{X_j^Z(s_1)X_l^Z(s_2)\}$ (resp. $C_{km}^Y(t_1,t_2) = \mathrm{E}\{Y_k^Z(t_1)Y_m^Z(t_2)\}$). Then, there exists an orthonormal eigenbasis, $\{\boldsymbol{\phi}_{Z,r}^X(s)\}_{r\ge1}$, where $\boldsymbol{\phi}_{Z,r}^X(s) = (\phi_{Z,1r}^X(s), \dots, \phi_{Z,pr}^X(s))^\top$ and $\langle \boldsymbol{\phi}_{Z,q}^X, \boldsymbol{\phi}_{Z,r}^X\rangle_{\mathbb{H}_1} = \delta_{rq}$, the Kronecker delta, with the corresponding non-negative eigenvalues $\{\lambda_{Z,r}^X\}_{r\ge1}$ in descending order, such that $\boldsymbol{C}^X(s_1,s_2) = \sum_{r=1}^\infty \lambda_{Z,r}^X\, \boldsymbol{\phi}_{Z,r}^X(s_1)\, \boldsymbol{\phi}_{Z,r}^X(s_2)^\top$, whose (j,l) element is $C_{jl}^X(s_1,s_2) = \sum_{r=1}^\infty \lambda_{Z,r}^X\, \phi_{Z,jr}^X(s_1)\, \phi_{Z,lr}^X(s_2)^\top$. Similarly, we define $\{\lambda_{Z,r}^Y\}_{r\ge1}$, and $\{\boldsymbol{\phi}_{Z,r}^Y(t)\}_{r\ge1}$ and $\boldsymbol{\phi}_{Z,r}^Y(t) = (\phi_{Z,1r}^Y(t), \dots, \phi_{Z,dr}^Y(t))^\top$ in $\mathbb{H}_2$ such that $\boldsymbol{C}^Y(t_1,t_2) = \sum_{q=1}^\infty \lambda_{Z,q}^Y\, \boldsymbol{\phi}_{Z,q}^Y(t_1)\, \boldsymbol{\phi}_{Z,q}^Y(t_2)^\top$.

2.2. Multivariate functional linear regression model

The multivariate functional linear regression (mFLR) model is based on the transformed variables $\boldsymbol{X}^Z$ and $\boldsymbol{Y}^Z$, which can be represented as follows:

$$\boldsymbol{X}^Z(s) = \sum_{r=1}^\infty \xi_{Z,r}^X \boldsymbol{\phi}_{Z,r}^X(s),\ s\in\mathcal{S} \quad \left(\text{resp. } \boldsymbol{Y}^Z(t) = \sum_{q=1}^\infty \xi_{Z,q}^Y \boldsymbol{\phi}_{Z,q}^Y(t),\ t\in\mathcal{T}\right), \tag{1}$$

where $\xi_{Z,r}^X = \langle \boldsymbol{X}^Z, \boldsymbol{\phi}_{Z,r}^X\rangle_{\mathbb{H}_1}$ (resp. $\xi_{Z,q}^Y = \langle \boldsymbol{Y}^Z, \boldsymbol{\phi}_{Z,q}^Y\rangle_{\mathbb{H}_2}$) is a random coefficient that satisfies $\mathrm{E}\xi_{Z,r}^X = 0$, $\mathrm{E}(\xi_{Z,r}^X\xi_{Z,q}^X) = \lambda_{Z,r}^X\delta_{rq}$ (resp. $\mathrm{E}\xi_{Z,q}^Y = 0$, and $\mathrm{E}(\xi_{Z,q}^Y\xi_{Z,r}^Y) = \lambda_{Z,q}^Y\delta_{qr}$). The components $\{X_l^Z\}$ in $\boldsymbol{X}^Z$ are correlated, so are the components $\{Y_l^Z\}$ in $\boldsymbol{Y}^Z$.