COMSTA: 6557 Model 3G pp. 1–16 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Time-dynamic varying coefficient models for longitudinal data

Kyeongeun Lee [a], Young K. Lee [b,*], Byeong U. Park [a], Seong J. Yang [c]

[a] Seoul National University, Republic of Korea
[b] Kangwon National University, Republic of Korea
[c] Hankuk University of Foreign Studies, Republic of Korea

ARTICLE INFO

Article history:
Received 3 August 2017
Received in revised form 16 January 2018
Accepted 28 January 2018
Available online xxxx

Keywords:
Kernel smoothing
Longitudinal data
Smooth backfitting
Varying coefficient models

ABSTRACT

A new varying coefficient model that relates functional response to functional predictors is proposed and studied. The model accommodates the influence of the functional predictors on the time-varying coefficient functions. A powerful kernel smoothing technique is developed for estimating the model with longitudinal observations of the functional response and predictors. The method involves a backfitting iteration that is based on alternating conditional expectation. The convergence of the algorithm is established and the asymptotic distribution of the coefficient function estimators is derived. It is shown that the method works well for finite sample sizes via simulation studies. The proposed model and method are also applied to analyzing an air quality dataset.

1. Introduction

There has been a vast amount of work on varying coefficient models. The readers are referred to Park et al. (2015) for a review of the earlier studies on the topic. Varying coefficient models are particularly useful in longitudinal analysis, where one is typically interested in investigating how the effect of predictors X_j on a response Y changes over time. The simplest functional regression model for this purpose is

$$Y(t) = \beta_1(t)X_1(t) + \cdots + \beta_d(t)X_d(t) + \epsilon(t), \tag{1.1}$$

where β_j are unknown coefficient functions and ϵ is a mean zero stochastic process. The model (1.1) is a direct extension of the classical linear model to the case of functional response and predictors. It has been studied by Hoover et al. (1998), Wu et al. (1998, 2000); Wu and Yu (2002), Huang et al. (2004), Şentürk and Müller (2008), Wang et al. (2008) and Noh and Park (2010), among others.

In the model (1.1) the coefficient functions β_j depend only on time t. This might be restrictive. The coefficient functions may change with the value of other functional predictors, say Z_j, $1 \le j \le d$. In this paper, we consider a new varying coefficient model that accommodates the latter situation. Specifically, we study the estimation of the model

$$Y(t) = \beta_1(t, Z_1(t))X_1(t) + \cdots + \beta_d(t, Z_d(t))X_d(t) + \epsilon(t), \tag{1.2}$$

where ϵ is a stochastic process such that $E(\epsilon|X_1, Z_1, \ldots, X_d, Z_d) = 0$. The model (1.2) is a functional version of the varying coefficient model for cross-sectional data,

$$Y = \beta_1(Z_1)X_1 + \cdots + \beta_d(Z_d)X_d + \epsilon, \tag{1.3}$$

* Correspondence to: Department of Statistics, Kangwon National University, Chuncheon 24341, Republic of Korea.
E-mail address: youngklee@kangwon.ac.kr (Y.K. Lee).

https://doi.org/10.1016/j.csda.2018.01.016
0167-9473/

Please cite this article in press as: Lee K., et al., Time-dynamic varying coefficient models for longitudinal data. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.01.016.

which was proposed by Hastie and Tibshirani (1993) and studied by Yang et al. (2006) and Lee et al. (2012a). The generalization is in the same spirit as that of the classical linear model to its functional version (1.1). Our model (1.2) adds possible dynamic features to the coefficients β_j in the simpler model

$$Y(t) = \beta_1(Z_1(t))X_1(t) + \cdots + \beta_d(Z_d(t))X_d(t) + \epsilon(t). \tag{1.4}$$

Note that in the latter model (1.4) the coefficients β_j accommodate the effects of time t only through the values of the predictors Z_j.

To the best of our knowledge, the estimation of the model (1.2) has not been studied before. Since the models (1.1) and (1.4) are submodels of (1.2), the estimators of β_j under the model (1.2) may be used for checking the validity of the models (1.1) and (1.4). Unlike the models (1.3) and (1.4) the coefficients β_j in the model (1.2) have a common component, time t, that affects their values, which brings about an additional complication in the estimation of the model. The problem of estimating the model (1.2) does not fit into the framework of estimating the additive model as in Mammen et al. (1999) or in its functional extension

$$Y(t) = \beta_1(t, Z_1(t)) + \cdots + \beta_d(t, Z_d(t)) + \epsilon(t) \tag{1.5}$$

studied by Zhang et al. (2013). Apparently, β_j in our model (1.2) are not additive components of the regression function as in (1.5). The model (1.2) takes into account nonlinear interaction effects between the two predictor groups, which is not the case with the model (1.5).

To mention a few other extensions of the linear model (1.1) for longitudinal data, Sun and Wu (2005) considered a semiparametric extension, and Şentürk and Müller (2010) studied a model where $X(t)$ is replaced by its history in a window $[t-\delta, t]$ in the form of $\int_0^\delta \gamma(u)X(t-u)\,du$ for some unknown function γ. Zhang and Wang (2015) introduced a varying coefficient model for time-invariant predictors, say Z_j, that replaces $X_j(t)$ in (1.1) by nonparametric functions of Z_j. As for a related work on the additive model (1.5), Ma and Zhu (2016) studied a different kind of additive models for functional data where the effects of a predictor process at different time points are integrated continuously across the time domain.

In this paper we present a powerful kernel smoothing technique that estimates the coefficient functions β_j in the model (1.2). Our estimators of $\beta_j(t, z)$ are smooth in both directions, t and z. The method is to solve a system of backfitting equations that results from alternating conditional expectation. We propose an iteration scheme to get the estimators of β_j as the solution of the system of backfitting equations. We prove that the iteration converges to the solution in an L_2 sense for the direction z along the direction t. We also derive the asymptotic distributions of the estimators of β_j, and show that the estimator of β_j, for each j, has the same asymptotic distribution of the oracle estimator that utilizes the knowledge of all other coefficients β_k for $k \neq j$. We present a simulation result that confirms the validity of the proposed method in finite sample sizes, and also illustrate the method through an air quality dataset.

2. Methodology

We study the estimation of the model (1.2) based on a longitudinal dataset for the response and predictor processes observed intermittently at discrete time points. We consider the case where the time points are allowed to be different for different subjects but are independent random variables governed by an unknown common distribution. In this longitudinal setting the observed response and predictors admit the following model

$$Y(T) = \beta_1(T, Z_1(T))X_1(T) + \cdots + \beta_d(T, Z_d(T))X_d(T) + \epsilon(T), \tag{2.1}$$

where T denotes the random variable that represents the independent time points where the processes $\mathbf{X} \equiv (X_1, \ldots, X_d)^\top$, $\mathbf{Z} \equiv (Z_1, \ldots, Z_d)^\top$ and Y are observed.

Our proposed method of estimating β_j in (1.2) is based on the local linear smoothing technique and alternating conditional expectation. To motivate the method, consider the univariate regression problem of estimating $m(u) = E(V|U = u)$ for the pair of a predictor U and a response V. The standard kernel estimator of $E(V|U = u)$ takes the form $\sum_{i=1}^n w_i(u)V_i$ with the 'local' weights $w_i(u) \geq 0$ determined by a baseline kernel function K in such a way that

$$w_i(u) = \left(n^{-1}\sum_{i=1}^n K_h(u, U_i)\right)^{-1} n^{-1}K_h(u, U_i), \tag{2.2}$$

where typically $K_h(u, U_i) = h^{-1}K((u - U_i)/h)$ and h is the bandwidth. The local linear estimator of $m(u)$ is defined along with an estimator of its derivative $m'(u)$. The estimator of $(m(u), m'(u))$ is usually defined in the literature as the minimizer of $\sum_{i=1}^n (V_i - m(u) - m'(u)(U_i - u))^2 K_h(u, U_i)$. But, we observe that it can be also defined as the solution of an empirical version of the following equation:

$$E\left(\left.\begin{pmatrix}1\\ U-u\end{pmatrix} V \,\right|\, U = u\right) = E\left(\left.\begin{pmatrix}1\\ U-u\end{pmatrix}(1, U-u) \,\right|\, U = u\right)\begin{pmatrix}m(u)\\ m'(u)\end{pmatrix}. \tag{2.3}$$