Journal of Computational and Applied Mathematics 300 (2016) 1–17

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Generalized varying index coefficient models

Chaohui Guo *, Hu Yang, Jing Lv

College of Mathematics and Statistics, Chongqing University, Chongqing, 401331, China

ARTICLE INFO

Article history:
Received 19 November 2014

Keywords:
Asymptotic normality
B splines
Generalized varying index coefficient model
Profile spline quasi-likelihood
Two-step spline backfitted local quasi-likelihood

ABSTRACT

In this paper, we propose a new semiparametric model called generalized varying index coefficient models (GVICMs). The GVICM is a generalization of the varying index coefficient model (VICM) proposed by Ma and Song (2014), by allowing for non-Gaussian data and nonlinear link functions. The GVICM serves as a good tool for modeling and assessing nonlinear interaction effects between grouped covariates and the response variable. Our main goal is to estimate the unknown parameters and nonparametric functions. Firstly, we develop a profile spline quasi-likelihood estimation procedure to estimate the regression parameters and nonparametric coefficients in which the nonparametric functions are approximated by *B*-spline basis functions. Under some mild conditions, we establish asymptotic normalities of parameter estimations as well as the convergence rates of nonparametric estimators. Secondly, we develop a two-step spline backfitted local quasi-likelihood estimation for achieving asymptotic distribution of nonparametric function. Moreover, the oracle property of the nonparametric estimator is also established by utilizing the two-step estimation approach. Simulation study and a set of real data are carried out to investigate the performance of the proposed method.

1. Introduction

Semiparametric models arise frequently in the field of statistics and practice. This is mainly because they retain the virtues of both parametric and nonparametric modeling, that is, they can reduce the high risk of misspecification with respect to fully parametric models and avoid some serious drawbacks such as the curse of dimensionality, difficulty of interpretation and lack of extrapolation capability relative to purely nonparametric methods. Due to these advantages mentioned above, semiparametric models have been paid more and more attention recently. Here we only list a few. See single index models [1,2], varying coefficient models [3,4], partially linear varying coefficient models [5,6], partially linear single index models [7,8], additive models [9,10], single-index varying coefficient models [11–14] and so on. Recently, Ma and Song [15] presented a new class of varying index coefficient models (VICMs) which are flexible by unifying diverse semiparametric regression models. The VICM has the following form

$$Y = m(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\beta}) + \varepsilon = \sum_{l=1}^{d} m_l(\boldsymbol{Z}^T \boldsymbol{\beta}_l) X_l + \varepsilon, \tag{1}$$

where $\boldsymbol{\beta}_l = (\beta_{l1}, \ldots, \beta_{lp})^T$ are the coefficient vectors which vary with different covariates X_l, and $m_l(\cdot)$ are unknown nonparametric function for $1 \le l \le d$. For the reason of identifiability, we generally assume $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_d^T\right)^T$ belongs to

* Corresponding author.
E-mail address: guochaohui2010@126.com (C. Guo).

http://dx.doi.org/10.1016/j.cam.2015.11.025
0377-0427/

the parameter space:

$$\Theta = \left\{ \boldsymbol{\beta} = \left(\boldsymbol{\beta}_l^T : 1 \le l \le d\right)^T : \|\boldsymbol{\beta}_l\| = 1, \beta_{l1} > 0, \boldsymbol{\beta}_l \in R^p \right\}, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. The VICM encompasses the above mentioned semiparametric models as special cases. The more details about the relationship of the VICM to the existing models can refer to the Section 2 of Ma and Song [15].

As far as we know, generalized linear models (GLMs) have more wide applications than linear models, since they can deal with different types of responses, for example, binary data and count data. Although the VICM is natural and useful modeling tool in many practical applications, we find that the VICM is only suitable for continuous data but not for discrete data. Motivated by the same spirit that generalized linear models (GLMs, McCullagh and Nelder [16]) provide an extension of linear models, we propose a new class of semiparametric models, namely, generalized varying index coefficient models (GVICM). The GVICM is a useful extension of the VICM in dealing with different types of responses. Let Y be a response variable and $(\boldsymbol{X}, \boldsymbol{Z})$ be its associated covariates, the GVICM takes the form

$$E(Y|\boldsymbol{X}, \boldsymbol{Z}) = \mu(\boldsymbol{X}, \boldsymbol{Z}) = g^{-1}(\eta), \quad \text{with } \eta = m(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\beta}) = \sum_{l=1}^{d} m_l(\boldsymbol{Z}^T\boldsymbol{\beta}_l)X_l, \tag{3}$$

where $g(\cdot)$ is a known link function, which is monotone and differentiable. Without loss of generality, we assume that the response variable of the GVICM follows an exponential family distribution, so we further assume that the conditional variance is a function of the mean defined by

$$\mathrm{Var}(Y|\boldsymbol{X}, \boldsymbol{Z}) = V(\mu(\boldsymbol{X}, \boldsymbol{Z})), \tag{4}$$

where V is a known positive function. The GVICM not only maintains all superior statistical properties of the VCIM but also possesses some properties that the VCIM do not have, for example, it can deal with the non-Gaussian data and nonlinear link functions. The proposed GVICM (3) is flexible, which includes various existing generalized semiparametric models. Here we only list a few. For example, when $d \equiv 1$ and $X_1 = 1$, it reduces to the generalized single index model [17]; when $m_l(\cdot)$ are set as constant for $l \ge 2$ and $X_1 = 1$, it reduces to the generalized partial linear single index model [18]; when $\boldsymbol{\beta}_l$ are known and $X_l \equiv 1$, it reduces to the generalized additive model [19–21]; when the link function g is identity, it reduces to the VICM. Therefore, the research regarding GVICM model is of great theoretical and practical significance. In this paper, our aim is to estimate the unknown coefficient vectors $\boldsymbol{\beta}_l$ and the nonparametric functions $m_l(\cdot)$ in the model (3) with a known link function $g(\cdot)$. As we all know, the computing speed of spline estimation approach is faster than that of kernel-based in semiparametric models (see [22,23]). Moreover, kernel-based methods may become very complicated to solve the problem of handling multiple nonparametric functions simultaneously. Thus, we firstly propose a profile quasi-likelihood estimation procedure based on B-spline basis functions approximations, and establish root n-consistency and asymptotic normality of the parameter vector $\boldsymbol{\beta}$. Then, a two-step spline backfitted local quasi-likelihood is developed for studying the asymptotic normality of the estimator of nonparametric function.

The rest of this article is organized as follows. In Section 2, we develop the profile spline quasi-likelihood approach for the GVICM, where the nonparametric functions are approximated by B-spline. Then asymptotic properties of the proposed estimators are presented in this section. In Section 3, in order to obtain the asymptotic distribution of the nonparametric function $m_l(\cdot)$, we describe a two-step spline backfitted local quasi-likelihood (SBLQL) estimation. In Section 4, we present an algorithm for the estimation procedure. Monte Carlo simulation studies and a real data analysis are used to illustrate the proposed methodology in Section 5. Some concluding remarks are given in Section 6. All the regularity conditions and the technical proofs are provided in the Appendix.

2. Profile spline quasi-likelihood estimator and sampling properties

2.1. Estimation procedure for the GVICM

Suppose that $\{(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i),\ 1 \le i \le n\}$ is an independent and identically distributed sample from the model (3), where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^T$ and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^T$. Our main interest is to estimate the coefficient vectors $\boldsymbol{\beta}_l$ and nonparametric functions $m_l(\cdot)$ for $l = 1, \ldots, d$. Then the conditional quasi-likelihood function is defined by

$$Q(\mu, y) = \int_y^{\mu} \frac{y - s}{V(s)} ds. \tag{5}$$

Denote by $L_n(\boldsymbol{\beta}, m)$ the quasi-likelihood of the collected data $\{(\boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i),\ 1 \le i \le n\}$. Thus, we can obtain the estimators of $\boldsymbol{\beta}$ and $m_l(\cdot)$ by maximizing the following quasi-likelihood function

$$L_n(\boldsymbol{\beta}, m) = \sum_{i=1}^{n} Q\left[g^{-1}\left\{\sum_{l=1}^{d} m_l(\boldsymbol{Z}_i^T\boldsymbol{\beta}_l)X_{il}\right\}, Y_i\right] \tag{6}$$