Journal of Multivariate Analysis 132 (2014) 94–110

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Penalized quadratic inference functions for semiparametric varying coefficient partially linear models with longitudinal data

Ruiqin Tian [a], Liugen Xue [a,*], Chunling Liu [b]

[a] College of Applied Sciences, Beijing University of Technology, Beijing, 100124, China

[b] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received 29 March 2012
Available online 8 August 2014

AMS subject classifications:
62G08
62G20

Keywords:
Semiparametric varying coefficient partially linear models
Variable selection
Longitudinal data
Quadratic inference functions

ABSTRACT

In this paper, we focus on the variable selection for semiparametric varying coefficient partially linear models with longitudinal data. A new variable selection procedure is proposed based on the combination of the basis function approximations and quadratic inference functions. The proposed procedure simultaneously selects significant variables in the parametric components and the nonparametric components. With appropriate selection of the tuning parameters, we establish the consistency and asymptotic normality of the resulting estimators. Extensive Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed variable selection procedure. We further illustrate the proposed procedure by an application.

1. Introduction

As introduced in [6,13], the varying coefficient model provides a natural and useful extension of the classical linear regression model by allowing the regression coefficients to depend on certain covariates. Due to its flexibility to explore the dynamic features which may exist in the data and its easy interpretation, the varying coefficient model has been widely applied in many scientific areas. It has also experienced rapid developments in both theory and methodology, see [11] for a comprehensive survey. Zhang et al. [30] noticed that in practice some of the coefficients are constant rather than varying and proposed the so-called semiparametric varying coefficient partially linear models. Statistically, treating constant coefficients as varying will degrade estimation efficiency. On the other hand, longitudinal data occur very frequently in biomedical studies. Qu et al. [21] proposed a method of quadratic inference functions (QIF). It avoids estimating the nuisance correlation structure parameters by assuming that the inverse of the working correlation matrix can be approximated by a linear combination of several known basis matrices. The QIF can efficiently take the within-cluster correlation into account and is more efficient than the GEE approach when the working correlation is misspecified. Qu and Li [20] applied the QIF method to varying coefficient models for longitudinal data. Bai et al. [3] extended the QIF method to the semiparametric partial linear model. Dziak et al. [7] gave an overview on QIF approaches for longitudinal data.

It is well known, variable selection is an important topic in all regression analyses, therefore, many procedures and criteria have been developed for this, especially for linear regression analysis. For example, some simple and commonly

* Corresponding author.
E-mail addresses: tianruiqin@hotmail.com (R. Tian), lgxue@bjut.edu.cn (L. Xue).

http://dx.doi.org/10.1016/j.jmva.2014.07.015

used approaches include the stepwise selection and subset selection. Other selection criteria include Akaike [2] information criterion (AIC), Mallows' [19] C_p and Bayesian information criterion (BIC; Schwarz [23]). Nevertheless, those selection methods suffer from expensive computational costs. As computational efficiency is more desirable in many situations, various shrinkage methods have been developed, which include but are not limited to: the nonnegative garrotte [5], the LASSO [24], the bridge regression [12], the SCAD [9], and the one-step sparse estimator [34]. Furthermore, a number of works have also been done to extend the shrinkage estimation methods to nonparametric models. For example, Fan and Li [10] proposed to use the SCAD penalty for variable selection in longitudinal data analysis. Li and Liang [17] carefully studied variable selection for varying coefficient partially linear models, where the parametric components are identified via the SCAD [9], but the nonparametric components are selected via a generalized likelihood ratio test instead of a shrinkage method. Recently, Zhao and Xue [31] proposed a variable selection method to select significant variables in the parametric components and the nonparametric components simultaneously. Zhao and Xue [32] developed the variable selection procedure by combining basis function approximations for semiparametric varying coefficient partially linear models when the covariates in the parametric and nonparametric components are all measured with errors.

Various methods are available for fitting the semiparametric varying coefficient partially linear models with longitudinal data, such as, the kernel smoothing method, empirical likelihood method and the penalized spline method. Ahmad et al. [1] proposed a general series method to estimate semiparametric varying coefficient partially linear models. Fan and Huang [8] developed a profile least-square technique for the estimation of the parametric component. Xue and Zhu [29] considered the empirical likelihood inference for a varying coefficient model with longitudinal data. In addition, Wang et al. [25] proposed a group SCAD procedure for variable selection of pure varying coefficient models with longitudinal data.

In this paper, we extend the group SCAD variable selection procedure to the semiparametric varying coefficient partially linear models with longitudinal data, and propose a variable selection procedure based on the basis function approximations with quadratic inference functions. Furthermore, with proper choice of tuning parameters, we show that this variable selection procedure is consistent, and the estimators of regression coefficients have oracle property. Here, the oracle property means that the estimators of the nonparametric components achieve the optimal convergence rate, and the estimators of the parametric components have the same asymptotic distribution as that based on the correct submodel.

We adopt the penalized QIF method for semiparametric varying coefficient partially linear models. Compared with Zhao and Xue [31,32], we considered the longitudinal data and incorporated the correlation structure. In contrast, although Xue et al. [28] considered the penalized QIF for the generalized additive model, the model they considered is just a special case of semiparametric varying coefficient model. In addition, they just selected the nonparametric function. However, our method can select significant variables in the parametric components and nonparametric components simultaneously.

The rest of this paper is organized as follows. In Section 2 we first propose a variable selection procedure for the semiparametric varying coefficient partially linear models with longitudinal data. Asymptotic properties of the resulting estimators are considered in Section 3. In Section 4 we give the computation of the penalized QIF estimator. In Section 5 we carry out simulation studies to assess the finite sample performance of the method. We further illustrate the proposed methodology via a real data analysis in Section 6. The article is concluded with a brief discussion in Section 7. Some assumptions and the technical proofs of all asymptotic results are provided in the Appendix.

2. Variable selection based quadratic inference functions

Consider a longitudinal study with n subjects and m_i observations over time for the ith subject ($i = 1, \ldots, n$) for a total of $N = \sum_{i=1}^{n} m_i$ observations. Each observation consists of a response variable Y_{ij} and the covariate vectors $X_{ij} \in R^p, Z_{ij} \in R^q, U_{ij} \in R$ taken from the i subject. We assume that the observations from different subjects are independent, but that those within the same subject are dependent. The semiparametric varying coefficient partially linear models with longitudinal data have the form

$$Y_{ij} = X_{ij}^T \beta + Z_{ij}^T \alpha(U_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m_i, \tag{2.1}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ vector of unknown regression coefficients, $\alpha(u) = (\alpha_1(u), \ldots, \alpha_q(u))^T$ is a $q \times 1$ vector of unknown functions. Here, we assume that u ranges over a nondegenerate compact interval, without loss of generality, that is assumed to be the unit interval [0, 1]. We further give assumptions on the first two moments of the observation $\{Y_{ij}\}$. Let $E(Y_{ij}) = \mu_{ij}$ and $\mathrm{Var}(Y_{ij}) = \upsilon(\mu_{ij})$, where $\upsilon(\cdot)$ is a known variance function.

Following He et al. [15], we replace $\alpha(\cdot)$ by its basis function approximations. More specifically, let $B(u) = (B_1(u), \ldots, B_L(u))^T$ be B-spline basis functions with the order of M, where $L = K + M$, and K is the number of interior knots. We use the B-spline basis functions because they have bounded support and are numerically stable [22]. The spline approach also treats a non-parametric function as a linear function with the basis functions as pseudodesign variables, and thus any computational algorithm developed for the generalized linear models can be used for the semiparametric varying coefficient partially linear models. Of course, the B-spline is not the only choice of the nonparametric approximation method here. For example, Qu and Li [20] applied the QIF approach and the penalized spline approximation method together with varying coefficient models for longitudinal data. Then, similar to He et al. [14], $\alpha_k(u)$ can be approximated by

$$\alpha_k(u) \approx B(u)^T \gamma_k, \quad k = 1, \ldots, q, \tag{2.2}$$

where γ_k is a $L \times 1$ vector of unknown regression coefficients.