Journal of Multivariate Analysis 130 (2014) 354–375

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Partially linear single index models for repeated measurements

Shujie Ma [a,*], Hua Liang [b], Chih-Ling Tsai [c]

[a] *Department of Statistics, University of California, Riverside, CA 92521, United States*

[b] *Department of Statistics, The George Washington University, 801 22nd St. NW, Washington, DC 20052, United States*

[c] *Graduate School of Management, University of California, Davis, CA 95616, United States*

ARTICLE INFO

Article history:
Received 20 December 2013
Available online 16 June 2014

AMS subject classifications:
primary 62G08
secondary 62G10
62G20
62J02
62F12

Keywords:
Consistency
Model selection
Oracle estimator
Polynomial spline
Profile principle
Quadratic inference function
SCAD

ABSTRACT

In this article, we study the estimations of partially linear single-index models (PLSiM) with repeated measurements. Specifically, we approximate the nonparametric function by the polynomial spline, and then employ the quadratic inference function (QIF) together with profile principle to derive the QIF-based estimators for the linear coefficients. The asymptotic normality of the resulting linear coefficient estimators and the optimal convergence rate of the nonparametric function estimate are established. In addition, we employ a penalized procedure to simultaneously select significant variables and estimate unknown parameters. The resulting penalized QIF estimators are shown to have the oracle property, and Monte Carlo studies support this finding. An empirical example is also presented to illustrate the usefulness of penalized estimators.

1. Introduction

In regression analysis with repeated measures over time (i.e., longitudinal/panel data or cluster data), semiparametric models have been used to take into account both linear and nonlinear effects of covariates. The pioneering work in this context can be traced back to Severini and Staniswalis [31], followed by a series of efforts, such as Chen et al. [6], Chen and Jin [8], Fan and Li [14], He et al. [17], Lin and Carroll [25], Lin and Carroll [24], and Su and Ullah [35]. All of these works focus on the case in which the models either contain one nonparametric component or a summand of nonparametric functions. The typical approach for estimating parameters in these models is kernel-based backfitting and local scoring, proposed by Buja et al. [4]. Although those models are very flexible, they have some limitations. For example, the model with one multivariate nonparametric function suffers from the "curse of dimensionality" when the number of covariates is moderate or large, while the model with a summand of nonparametric functions does not take into account the interaction effects among covariates. In addition, the backfitting estimation algorithm can become computationally expensive when the number of covariates is large. This is because it requires extensive iterations to update the estimator of each nonparametric component [16,44]. This motivates us to adapt partially linear single-index models (PLSiM) to analyze repeatedly measured

[*] Corresponding author.
E-mail addresses: shujie.ma@ucr.edu (S. Ma), hliang@gwu.edu (H. Liang), cltsai@ucdavis.edu (C.-L. Tsai).

http://dx.doi.org/10.1016/j.jmva.2014.06.011
0047-259X/© 2014 Elsevier Inc. All rights reserved.

data. Semiparametric single-index models have been widely used as an appealing and effective statistical tool to model the relationship between the response variable and multivariate covariates, since it achieves dimension reduction and relaxes the restrictive parametric assumptions. See [18] for detailed discussion and illustration of the usefulness of this model. The PLSiM as a natural extension allows discrete explanatory variables to be modeled in the linear part. See [23,41,5,7] for studies and applications of PLSiM.

To estimate parameters in PLSiM, various methods have been proposed, including the backfitting algorithm [5], the penalized spline [45], the average derivative estimation method (ADE, [27]), the minimum average variance estimation (MAVE, [42,40]), the profile least squares approach (PrLS, [20,22]), and the estimating function method (EFM, [9]). It is noteworthy that the backfitting algorithm can be unstable and penalized spline estimation may not be efficient. Although the ADE, MAVE and profile methods overcome these limitations, ADE may suffer from the curse of dimensionality as mentioned in [39], MAVE may encounter the sparseness problem as noted by Cui et al. [9], and the PrLS estimator is not easy to obtain due to minimizing a high-dimensional nonlinear objective function. In addition, the above methods mainly focus on cross-sectional data rather than repeatedly measured data. Moreover, the true correlation structure within each cluster is often unknown, and ignoring such a correlation could yield biased estimators [37], inefficient estimators, and low power in hypothesis testing. Hence, we need to use a sophisticated method to parsimoniously separate within-subject and between-subject variation, and should not simply treat repeated measurements as cross-sectional observations. Consequently, developing an effective estimation procedure and then establishing its theoretical justifications for PLSiM with repeatedly measured data become an important and challenging task.

To alleviate the impact of correlation misspecification and to pursue estimation efficiency, we employ the quadratic inference function (QIF) proposed by Qu and Lindsay [29], Qu et al. [30], used for estimation in parametric models. This approach enables us to take into account the within-cluster correlation without specifying the covariance function. Furthermore, it is more efficient than the generalized estimating equation (GEE) approach when the working correlation is misspecified, as demonstrated in [30]. In this paper, we propose a QIF estimation procedure by incorporating the profile principle [32] for PLSiM with repeated measurements. Specifically, it consists of two steps: (i) For the given parametric components, employ the QIF approach to obtain an estimate of the nonparametric component by polynomial splines—It is noteworthy that the resulting nonparametric estimator is a function of the given parametric components; (ii) Based on the nonparametric estimator, construct the profiled QIF objective function for the parametric components and then obtain their estimators.

The QIF approach has been recently applied to estimation in single-index models [1] and PLSiM [21] with the nonparametric functions estimated by penalized-splines and local linear smoothing, respectively. The spline estimation approach is known as computationally faster and more efficient than kernel smoothing in semiparametric models with correlated data [26]. It is noteworthy that Bai et al. [1] studied the asymptotic normality for the index parameters by assuming that the true nonparametric link function is known. Comparing to Bai et al. [1], we derive root-n consistency and a sandwich formula for the covariance matrix of the parametric estimators by estimating the link function with polynomial splines. Therefore, to obtain the asymptotic properties, we face significant theoretical challenges since the parametric QIF estimators are involved in the nonparametric functional estimator with divergent parameters. Thus, the classical asymptotic theory cannot be directly applied. Accordingly, we explore a new approach to establish the asymptotic normality of the parametric estimators in the PLSiM. Another contribution in our paper is that we introduce the penalized QIF (PQIF) to reduce the model complexity, which shrinks irrelevant coefficients of the linear and single-index components to zero. The resulting estimators of the nonzero coefficients are shown to be asymptotically normal and have the oracle property. Furthermore, Xue et al. [43] applied the QIF to additive models and studied the optimal convergence rate of the spline estimators for the additive nonparametric functions.

The paper is organized as follows. Section 2 introduces the models and then applies QIF to obtain parametric and nonparametric estimations. The theoretical properties of parametric estimators are established. Section 3 proposes a penalized quadratic inference function approach for PLSiM to simultaneously estimate parameters and select variables, and the resulting estimators possess the oracle property. The practical implementations are developed in Section 4, and simulation studies and an empirical example are presented in Section 5. The last section concludes the article with a brief discussion, and technical proofs are given in the Appendix.

2. Models and estimation methods

2.1. Models

Suppose that the data consist of n independent subjects and the ith ($i = 1, \dots, n$) cluster has m_i repeated measures. Let Y_{ij} be the response variable, and $X_{ij} = \left(X_{ij,1}, \dots, X_{ij,d_1}\right)^{\mathrm{T}}$ and $Z_{ij} = \left(Z_{ij,1}, \dots, Z_{ij,d_2}\right)^{\mathrm{T}}$ be $d_1 \times 1$ and $d_2 \times 1$ covariate vectors, respectively, for the jth observation in the ith cluster. Denote $Y_i = \left(Y_{i1}, \dots, Y_{im_i}\right)^{\mathrm{T}}$, $\boldsymbol{X}_i = \left(X_{i1}, \dots, X_{im_i}\right)^{\mathrm{T}}_{m_i \times d_1}$, and $\boldsymbol{Z}_i = \left(Z_{i1}, \dots, Z_{im_i}\right)^{\mathrm{T}}_{m_i \times d_2}$. Consider the partially linear single-index model (PLSiM),

$$E\left(Y_{ij} \left| X_{ij}, Z_{ij}\right.\right) = \mu_{ij} = g\left(\boldsymbol{\beta}^{\mathrm{T}} X_{ij}\right) + \boldsymbol{\alpha}^{\mathrm{T}} Z_{ij}, \tag{2.1}$$