Journal of the Korean Statistical Society 43 (2014) 303–314

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Variable selection in robust semiparametric modeling for longitudinal data

Kangning Wang [a,b], Lu Lin [a,*]

[a] Shandong University Qilu Securities Institute for Financial Studies and School of Mathematics, Shandong University, Jinan, China
[b] Department of Mathematics & KLDAIP, Chongqing University of Arts and Sciences, Chongqing, China

ARTICLE INFO

Article history:
Received 4 June 2013
Accepted 23 October 2013
Available online 11 November 2013

AMS 2000 subject classifications:
primary 62G05
secondary 62E20

Keywords:
Semiparametric model
Longitudinal data
Robustness
M-type estimator
Variable selection
Oracle property

ABSTRACT

This paper considers robust variable selection in semiparametric modeling for longitudinal data with an unspecified dependence structure. First, by basis spline approximation and using a general formulation to treat mean, median, quantile and robust mean regressions in one setting, we propose a weighted M-type regression estimator, which achieves robustness against outliers in both the response and covariates directions, and can accommodate heterogeneity, and the asymptotic properties are also established. Furthermore, a penalized weighted M-type estimator is proposed, which can do estimation and select relevant nonparametric and parametric components simultaneously, and robustly. Without any specification of error distribution and intra-subject dependence structure, the variable selection method works beautifully, including consistency in variable selection and oracle property in estimation. Simulation studies also confirm our method and theories.

1. Introduction

Semiparametric models are often considered for analyzing longitudinal data for a good balance between flexibility and parsimony. Suppose that we have n subjects, among which the ith subject has $m_i \geq 1$ repeated measurements. Consider the following partially linear varying coefficient (PLVC) model for longitudinal data:

$$y_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\alpha}(t_{ij}) + \boldsymbol{z}_{ij}^{\top}\boldsymbol{\beta} + \epsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, m_i, \tag{1.1}$$

where y_{ij} denotes the jth outcome of the ith subject, $\boldsymbol{\alpha}(t) = (\alpha_1(t), \ldots, \alpha_p(t))^{\top} \in \mathbb{R}^p$ for $t \in [0, 1]$ is unknown but smooth function vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^{\top} \in \mathbb{R}^q$ is the constant coefficient vector, whose true values are $\boldsymbol{\alpha}_0(t)$ and $\boldsymbol{\beta}_0$ respectively, $\boldsymbol{x}_{ij} = (x_{ij}^1, \ldots, x_{ij}^p)^{\top} \in \mathbb{R}^p$ and $\boldsymbol{z}_{ij} = (z_{ij}^1, \ldots, z_{ij}^q)^{\top} \in \mathbb{R}^q$ are the design vectors. We assume that the observations, and therefore ϵ_{ij}, are dependent within the same subjects, but independent across subjects, and the form of the error distribution and the intra-subject dependence structure are left unspecified. The aim of this work is to select significant variables in the parametric and nonparametric components simultaneously and robustly.

Shrinkage-type variable selection methods have seen increasing applications, and numbers of works have been done to extend this type methods to varying coefficient (VC) or the PLVC models. Li and Liang (2008) studied variable selection for the PLVC models, where the parameter components are identified via the SCAD (Fan & Li, 2001) but the nonparametric components are selected via a generalized likelihood test, instead of the shrinkage method. Wang, Li, and Huang (2008) and Wang and Xia (2009) proposed two variable selection methods for the pure VC model through basis spline approximation

* Corresponding author. Tel.: +86 531 88364791.
E-mail address: linlu@sdu.edu.cn (L. Lin).

1226-3192/$ – see front matter
http://dx.doi.org/10.1016/j.jkss.2013.10.003

and kernel smoothing, respectively. Zhao and Xue (2009) presented a method via the SCAD which can simultaneously select parametric and nonparametric components in (1.1).

Methods in the aforementioned papers are mainly built on conditional mean regression. However, their performance can be adversely influenced by outliers in either the response or the covariate space. In this work, we first propose a weighted M-type regression, which utilizes a weight function to downweight the effect of leverage points and uses a general M-type loss function that treats mean, median, quantile and robust mean regressions in one setting. Thus the new method can keep balance between robustness and efficiency, and accommodate heterogeneity by choosing appropriate M-type loss function. Under mild conditions, the convergence rate of the estimator for $\boldsymbol{\alpha}(t)$ is obtained and the estimator for the $\boldsymbol{\beta}$ is shown to be asymptotically normally distributed. Furthermore, we propose a partial adaptive group L_2 norm penalized weighted M-type regression estimator, which can do estimation and select relevant nonparametric and parametric components simultaneously and robustly. Theoretical results show, with proper choice of tuning parameters, variable selection is consistent, and estimators enjoy the 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 obtained under the true model.

It is remarkable that combining the penalization with M-type regression is not trivial, especially in the semiparametric models. Notice that the loss function used in M-type regression may not be differentiable at some points (e.g., quantile regression loss); as a result, the general asymptotic results for penalized mean regression (e.g. Wang et al., 2008, Wang & Xia, 2009, Zhao & Xue, 2009) do not apply directly. Furthermore, in our semiparametric setting, in order to establish the asymptotic results, it involves three types of regularization parameters, i.e., the smoothing parameters (e.g., knot number), tuning parameters for parametric part and tuning parameters for nonparametric part. Due to their interaction, what should be the right convergence speed for the regularization parameters under this situation is much less well understood. On the other hand, since quantile regression loss involves a non-differentiable loss function that can be considered as an asymmetric L_1 function, the computation is challenging when the partial adaptive group L_2 norm penalty is used.

Recently, there are much development in variable selection in nonparametric and semiparametric models. Lin and Zhang (2006) proposed the component selection and smoothing operator (COSSO) method for an additive model. Huang, Horowitz, and Wei (2010) studied the LASSO (Tibshirani, 1996; Zou, 2006) for variable selection in the additive model. Wei, Huang, and Li (2011) studied variable selection issue for a high-dimensional varying coefficient model. Ma and Du (2012) studied variable selection in partial linear regression with diverging dimensions for right censored data. Zhang, Cheng, and Liu (2011) proposed a method for determining the zero, linear and nonlinear components in partially linear models. Huang, Wei, and Ma (2012) further proposed a novel semiparametric model pursuit method for identifying the covariates with linear effects and those with nonlinear effects, and proved selection consistency. Huang, Breheny, and Ma (2012) gave an excellent discussion of group selection. Furthermore, much research has been done based on the M-type regression. See, e.g. He and Shi (1994) and He, Zhu, and Fung (2002) for M-type regression estimation for partially linear models; Kim (2007) and Wang, Zhu, and Zhou (2009) for quantile regression estimation and assessment. For variable selection, also there are much development. Wu and Liu (2009) discussed variable selection for linear model by quantile regression; Hohsuk, Kwanghun, and Ingrid (2012), Tang, Wang, and Zhu (2013), Tang, Wang, Zhu, and Song (2012) and Zhao, Zhang, Lv, and Liu (2012) discussed variable selection issues by quantile regression for the pure VC model; Kai, Li, and Zou (2011) discussed variable selection only for the parametric component in the PLVC model by quantile regression; Li, Peng, and Zhu (2011) propose a nonconcave penalized M-estimation for the linear model and established the oracle property; Zhou, Jiang, and Qian (2013) proposed a least absolute deviations (LAD) variable selection for linear models with randomly censored data through the LASSO; Yao and Wang (2013) developed a robust sparse MAVE (Xia, Tong, Li, & Zhu, 2002) based on M-estimation.

The rest of this paper is organized as follows. In Section 2, we introduce our method and investigate its theoretical properties. The implementation issues are discussed in Section 3. Numerical studies are reported in Section 4. All the technical proofs are provided in the Appendix.

2. Methodology and asymptotic properties

2.1. The methodology

Let $\boldsymbol{\pi}(t) = \left(B_1(t), \ldots, B_{q_n}(t)\right)^\top$ be a set of B-spline basis functions of order $\hbar+1$ with K_n internal knots and $q_n = K_n+\hbar+1$. Then $\alpha_l(t)$ can be approximated as

$$\alpha_l(t) \approx \sum_{s=1}^{q_n} B_s(t)\theta_{l,s} = \boldsymbol{\pi}(t)^\top \boldsymbol{\theta}_l, \tag{2.1}$$

where $\{\boldsymbol{\theta}_l = \left(\theta_{l,1}, \ldots, \theta_{l,q_n}\right)^\top \in \mathbb{R}^{q_n}\}_{l=1}^p$ are spline coefficient vectors. Then, model (1.1) can be approximated as

$$y_{ij} \approx \sum_{l=1}^{p}\sum_{s=1}^{q_n} x_{ij}^l B_s(t_{ij})\theta_{l,s} + \sum_{k=1}^{q} z_{ij}^k \beta_k + \epsilon_{ij} = \boldsymbol{\Pi}_{ij}^\top \boldsymbol{\Theta} + \boldsymbol{z}_{ij}^\top \boldsymbol{\beta} + \epsilon_{ij}, \tag{2.2}$$

where $\boldsymbol{\Pi}_{ij} = \left(x_{ij}^1 \boldsymbol{\pi}_{ij}^\top, \ldots, x_{ij}^p \boldsymbol{\pi}_{ij}^\top\right)^\top \in \mathbb{R}^{pq_n}$, $\boldsymbol{\Theta} = \left(\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_p^\top\right)^\top \in \mathbb{R}^{pq_n}$ and $\boldsymbol{\pi}_{ij} = \boldsymbol{\pi}(t_{ij})$.