Journal of Computational and Applied Mathematics 280 (2015) 125–140

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Robust smooth-threshold estimating equations for generalized varying-coefficient partially linear models based on exponential score function

Jing Lv *, Hu Yang, Chaohui Guo

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

ARTICLE INFO

Article history:
Received 11 June 2014

Keywords:
Estimating equations
Generalized varying-coefficient partially linear models
Robust estimation
Oracle property

ABSTRACT

In this paper, we present a new efficient and robust estimation procedure for generalized varying coefficient partially linear models (GVCPLMs), where the nonparametric coefficients are approximated by polynomial splines. A bounded exponential score function with a tuning parameter γ and leverage based weights are applied to the estimating equations for achieving robustness against outliers in both the response and covariates directions. Our motivation for the new estimation procedure is that it enables us to achieve better robustness and efficiency by selecting automatically the tuning parameter γ using the observed data. The proposed estimator is as asymptotically efficient as the common quasi-likelihood estimator when there are no outliers. Moreover, an automatic variable selection procedure is developed to select significant parametric components for the GVCPLM based on robust smooth-threshold estimating equations. Simulations and a real data example are used to demonstrate the finite sample behavior of the proposed estimator.

1. Introduction

Generalized varying-coefficient partially linear models [1] extend the framework of varying coefficient partially linear models (VCPLMs), by allowing for non-Gaussian data and nonlinear link functions. This is same spirit as generalized linear models (GLMs, [2]) provide an extension of linear models. GVCPLMs have become a favored tool for the modeling of various responses, especially for binomial or Poisson type responses. More specifically, let Y be a real-valued response variable which may be continuous or discrete and $(U, \mathbf{X}, \mathbf{Z})$ be the associated covariates. We assume

$$g\{E(Y|U=u, \boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z})\} = g\{\mu_0(u, \mathbf{x}, \mathbf{z})\} = \boldsymbol{x}^T\boldsymbol{\beta}_0 + \boldsymbol{z}^T\boldsymbol{\alpha}_0(u), \tag{1}$$

where $g(\cdot)$ is a known link function, which is monotone and differentiable, $\boldsymbol{\beta}_0$ are the coefficients in the linear part and $\boldsymbol{\alpha}_0(u)$ is a vector consisting of unknown smooth regression coefficient functions. Corresponding covariates are denoted by $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ and $\boldsymbol{z} = (z_1, \ldots, z_q)^T$. $U \in \mathbb{R}$ is the univariate index variable such as the measurement time and we assume $U \in [0, 1]$ for simplicity.

In this paper, we only focus on that the response variable of GVCPLMs follows an exponential family distribution, so we further assume that the conditional variance is a function of the mean defined by

$$Var(Y|U=u, \boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Z}=\boldsymbol{z}) = V(\mu_0) = V(g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}_0 + \boldsymbol{z}^T\boldsymbol{\alpha}_0(u))), \tag{2}$$

* Corresponding author.
E-mail address: cqulv2015@126.com (J. Lv).

http://dx.doi.org/10.1016/j.cam.2014.11.003
0377-0427/

where V is a known function. The GVCPLM maintains the flexibility and explanatory power of generalized nonparametric regression models and generalized linear regression models respectively.

For the GVCPLM, many popular procedures for the estimation and inference have been well studied including [3–7]. However, all the above mentioned papers were built on quasi-likelihood method, which are expected to be very sensitive to outliers and their efficiency may be significantly reduced for contaminated data. Hence, robustness against outliers is a very important issue in the GVCPLM. The traditional robust M-estimations (a popular choice is Huber's estimation) have attracted much attention and have been discussed in many literatures. For example, [8] considered robust inference for GLMs; [9] proposed robust quasi-likelihood equations to obtain functional estimates for generalized additive models; Wong et al. [10] studied M-type estimators for fitting robust generalized additive models in the presence of anomalous data. These articles above all adopted Huber's score function on the Pearson residuals to dampen the effect of outliers in the response.

As far as we know, although Huber's score function is a robust modeling tool, it has limitation in terms of efficiency of estimation. Hence, looking for other bounded score functions with a superior efficiency is of great interest. Recently, Zhang et al. [11] and Zhao et al. [12] proposed a new efficient and robust estimation procedure called modal regression for the VCPLM by adopting the normal density on the residuals. The outstanding merit of their procedure is that it can obtain both robustness and efficiency by introducing an additional tuning parameter. Wang et al. [13] adopted a similar idea and developed a class of penalized robust regression estimators by using exponential squared loss. For example, for the linear regression model $Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_0 + \varepsilon_i$, Wang et al. [13] proposed to estimate the regression parameter $\boldsymbol{\beta}_0$ by minimizing

$$Q_\gamma(\boldsymbol{\beta}) = \sum_{i=1}^{n} (1 - \varphi_\gamma(t_i)), \tag{3}$$

where $\varphi_\gamma(t_i) = \exp\left(-t_i^2/\gamma\right)$ with $t_i = Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta}$, $\gamma > 0$ determines the degree of robustness of the estimation. The more descriptions about γ how to control the robustness and efficiency can refer to [13]. Obviously, minimizing the objective function (3) is equivalent to solve the following estimating equations

$$\sum_{i=1}^{n} \boldsymbol{X}_i \psi_\gamma(t_i) = \boldsymbol{0} \tag{4}$$

where $\psi_\gamma(t) = \dot{\varphi}_\gamma(t) = -\frac{2t}{\gamma} \exp\left(-t^2/\gamma\right)$, $\psi_\gamma(\cdot)$ is the first derivative of $\varphi_\gamma(\cdot)$. It is not difficult to verify that $\psi_\gamma(t)$ is also a bounded score function since $\psi_\gamma(t)$ will go to zero when t approaches infinity. The resulting estimators of solving Eq. (4) is more efficient than existing robust methods and is as asymptotically efficient as the least-squares-based estimator when there are no outliers. This fact motivates us to extend the estimating equation (4) to the GVCPLM. Thus our work is a natural extension of Zhao et al. [12] to more general types of responses by using robust estimating equations as opposed to VCPLMs with modal regression.

In many practical situations, some covariates in the parametric part of model (1) are inactive. To enhanced model predictability and interpretability, we consider sparse GVCPLMs. In recent years, penalization or shrinkage based variable selection methods have been developed to carry out variable selection including SCAD [14], the elastic net [15], the adaptive lasso [16] and the adaptive elastic net [17]. In this paper, we develop an effective and robust variable selection procedure to select significant parametric components in model (1) by using the smooth-threshold estimating equations proposed in [18]. Compared with shrinkage methods mentioned above, our approach can avoid convex optimization and is thus computationally simpler.

The rest of the paper is organized as follows. In Section 2, we present our estimation approach for GVCPLMs and describe asymptotic properties of the proposed method. In Section 3, we develop an efficient and robust variable selection procedure based on smooth-threshold estimating equations to automatically find the relevant parameters among all candidate parameters and estimate them simultaneously. Its oracle property is also established in this Section. In Section 4, an iterative algorithm is developed to implement the procedures. In addition, we discuss how to select the tuning parameter γ and other regularization parameters so that the corresponding estimators of the newly method are robust and sparse. In Section 5, we report some simulation studies and a real data example to examine the finite sample performance. Some concluding remarks are given in Section 6. Proofs of the theorems are provided in the Appendix.

2. Methodology

2.1. Robust estimating equations

Suppose that $\{U_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, Y_i, i = 1, \ldots, n\}$ is an independent and identically distributed sample. Let $\boldsymbol{B}(u) = (B_1(u), \ldots, B_K(u))^T$, we further assume that each nonparametric component $\alpha_j(u), j = 1, \ldots, q$, can be approximated by B-spline functions, that is

$$\alpha_j(u) \approx \sum_{k=1}^{K} a_{jk} B_k(u), \quad j = 1, \ldots, q \tag{5}$$

where $K = k_n + m + 1$ is the number of basis functions, m is the degree of the spline and k_n is the number of internal knot which controls the smoothness of the nonparametric coefficients. For simplicity, in our paper, we use same K (the number of basis functions) to approximate different nonparametric components and equally spaced knots are used here.