Statistics and Probability Letters 89 (2014) 58–64

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

New efficient estimation and variable selection in models with single-index structure

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

[a] Department of Mathematics & KLDAIP, Chongqing University of Arts and Sciences, Chongqing, China

[b] Shandong University Qilu Securities Institute for Financial Studies and School of Mathematics, Shandong University, Jinan, China

ARTICLE INFO

Article history:
Received 2 September 2013
Received in revised form 10 January 2014
Accepted 17 February 2014
Available online 6 March 2014

MSC:
62G05
62E20
62J02

Keywords:
Linearity condition
Composite quantile regression
Adaptive LASSO
Oracle property
Single-index structure

ABSTRACT

We propose a new efficient root-n consistent estimate for the direction of the index parameter vector in a class of models with single-index structure. To select the important predictors, we suggest using the adaptive LASSO and establish the oracle property. Simulation results also confirm the theoretical findings.

1. Introduction

Let y be a response variable and $\boldsymbol{x} = (x_1, \ldots, x_d)^\tau$ be a covariate vector. In most applications, to mitigate the risk of model misspecification and to overcome the "curse of dimensionality", semiparametric models have attracted much attention. Popular models include the response transformation model: $g_1(y) = \boldsymbol{\theta}_0^\tau \boldsymbol{x} + \epsilon$ and the classical single-index model: $y = g_2(\boldsymbol{\theta}_0^\tau \boldsymbol{x}) + \epsilon$. Here $g_1(\cdot)$ is an unknown monotone function, $g_2(\cdot)$ is an unknown link function, and ϵ is assumed to be independent of $\boldsymbol{x}$. For more details, see Horowitz (1996). A common feature in the model structure about these two classes of models is that the information of the response can be captured through a single linear combination of the covariates called the single-index structure.

In this paper, we consider the following class of models with a single-index structure

$$y = G(\boldsymbol{\theta}_0^\tau \boldsymbol{x}, \epsilon), \tag{1.1}$$

where $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,d})^\tau$, or equivalently

$$y \perp \boldsymbol{x} | \boldsymbol{\theta}_0^\tau \boldsymbol{x}, \tag{1.2}$$

which was originally proposed in Li and Duan (1989) and Li (1991), where $G(\cdot)$ is the unspecified link function, and $\perp$ indicates independence. The semiparametric model (1.1) is equivalent to saying that the response y is independent of $\boldsymbol{x}$ given the

[*] Corresponding author at: Department of Mathematics & KLDAIP, Chongqing University of Arts and Sciences, Chongqing, China. Tel.: +86 18765890870. *E-mail addresses:* wkn1986@126.com, wknsuda@126.com (K. Wang).

http://dx.doi.org/10.1016/j.spl.2014.02.019
0167-7152/© 2014 Elsevier B.V. All rights reserved.

index $\boldsymbol{\theta}_0^\tau \boldsymbol{x}$. Many important models, including linear models, the transformation linear model and the classical single-index model mentioned above, naturally satisfy (1.2). Obviously, when $G(\cdot)$ is not specified, the vector $\boldsymbol{\theta}_0$ is identifiable only up to a multiplicative scalar because any location-scale change in $\boldsymbol{\theta}_0^\tau \boldsymbol{x}$ can be absorbed into the link function. Thus we are only concerned with the direction of $\boldsymbol{\theta}_0$. With a known direction of $\boldsymbol{\theta}_0$, the scatter plot of y versus $\boldsymbol{\theta}_0^\tau \boldsymbol{x}$ suffices to provide information about $G(\cdot)$ (Cook, 1998). Thus, regardless of the link function $G(\cdot)$, our aim is to provide an efficient and consistent estimator for the direction of $\boldsymbol{\theta}_0$, and it suffices to produce a sparse estimate of the direction of $\boldsymbol{\theta}_0$ for variable selection.

Models (1.2) have been well-studied in the literature. For estimating the index $\boldsymbol{\theta}_0$, we include but are not limited to the least squares (LS) method (Li and Duan, 1989), the quantile regression based method for the classical single-index model (Zhu et al., 2012; Fan and Zhu, 2013), the structural adaptation method (Dalalyan et al., 2008; Hristache et al., 2001), and those in the sufficient dimension reduction context, such as (Li, 1991; Cook and Weisberg, 1991; Cook and Ni, 2005; Li and Wang, 2007; Zhu et al., 2011). On the other hand, some attempts have also been made to address the variable selection problem. Kong and Xia (2007) proposed new selection criteria for variable selection in the classical single-index model. Wu and Li (2011) investigated the asymptotic properties of sufficient dimension reduction estimators equipped with SCAD penalty (Fan and Li, 2001). Peng and Huang (2011) proposed a penalized LS for the classical single index model. Wang et al. (2012) and Zhu et al. (2011) proposed variable selection methods for (1.2). However, these methods are mainly built upon LS, which are very sensitive to the outliers and their efficiency may be dramatically reduced for heavy-tail error distribution. Therefore, it is of great interest to see whether the robust and efficient methods, such as those in Zou and Yuan (2008), have their justifiable counterparts in the general setting of (1.2) where no parametric model is imposed.

In this paper, we will propose the new estimation and variable selection method for model (1.2). We first estimate the direction of $\boldsymbol{\theta}_0$ using composite quantile regression (CQR) (Zou and Yuan, 2008). We show that the resulting estimate is consistent. We then employ the adaptive LASSO (Zou, 2006) to do variable selection. This paper makes the following contributions and uniqueness to the literature.

(i) Under mild conditions, we show that for any link function $G(\cdot)$ and distribution for the error, the simple linear CQR coefficient for $y|\boldsymbol{x}$ is proportional to $\boldsymbol{\theta}_0$ in model (1.2).

(ii) We prove that the resulting adaptive LASSO penalized simple linear CQR estimate for model (1.2) enjoys the oracle property.

The theoretical result in (i) implies that the ordinary linear CQR actually results in a root n consistent estimate for the direction of $\boldsymbol{\theta}_0$, does not need to estimate the involved nonparametric transformation or link function and has no distribution constraint for the error. This will bring much convenience for calculation; furthermore, the CQR can provide an effective and robust modeling tool for model (1.2). Result in (ii) indicates that the variable selection procedure works very well as if the true relevant variables in model (1.2) were known in advance.

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

2. The new method

2.1. Direction recovery and estimation

In this section, we first propose a method to recover the direction of $\boldsymbol{\theta}_0$ in the population level, and then discuss the asymptotic properties of the estimation in the sample level. Denote $0 < \tau_1 < \tau_2 < \cdots < \tau_K < 1$. Let

$$\mathcal{L}(b_1, \ldots, b_K, \boldsymbol{\theta}) := \sum_{k=1}^{K} E\left\{\rho_{\tau_k}\left(y - b_k - \boldsymbol{\theta}^\tau \boldsymbol{x}\right)\right\}, \tag{2.1}$$

where $\rho_{\tau_k}(u) = \tau_k u - u\mathbf{1}_{(u<0)}$, and $\tau_k = \frac{k}{1+K}$, $k = 1, \ldots, K$. Define

$$\left(\check{b}_1, \ldots, \check{b}_K, \check{\boldsymbol{\theta}}\right) := \underset{b_1,\ldots,b_K,\boldsymbol{\theta}}{\arg\min} \left\{\mathcal{L}(b_1, \ldots, b_K, \boldsymbol{\theta})\right\}, \tag{2.2}$$

here, b_ks are used to estimate the τ_k-quantiles of $y - \check{\boldsymbol{\theta}}^\tau \boldsymbol{x}$, and it is easy to verify that $\check{b}_k$ is the τ_k-th quantile $y - \check{\boldsymbol{\theta}}^\tau \boldsymbol{x}$.

Theorem 1. *If the covariate* $\boldsymbol{x}$ *in* (1.2) *satisfies the following linear condition*

$$E(\boldsymbol{x}|\boldsymbol{\theta}_0^\tau \boldsymbol{x}) = \mathrm{var}(\boldsymbol{x})\boldsymbol{\theta}_0 \left\{\boldsymbol{\theta}_0^\tau \mathrm{var}(\boldsymbol{x})\boldsymbol{\theta}_0\right\}^{-1} \boldsymbol{\theta}_0^\tau \boldsymbol{x}, \tag{2.3}$$

then $\check{\boldsymbol{\theta}} = \kappa \boldsymbol{\theta}_0$, *for some constant* κ.

Remark 1. Theorem 1 implies that the simple linear CQR estimation can offer a consistent estimator for the direction of the index vector $\boldsymbol{\theta}_0$ under the linear conditions (2.3), which is widely assumed in the context of sufficient dimension reduction. Li (1991) pointed out that the (2.3) is satisfied when $\boldsymbol{x}$ follows an elliptical distribution. Hall and Li (1993) proved that this linearity condition always holds to a good approximation in single-index models of the form (1.1) when the dimension d diverges. Thus, the linearity condition is typically regarded as mild, particularly when d is fairly large.