Journal of the Korean Statistical Society 44 (2015) 606–618

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Robust direction identification and variable selection in high dimensional general single-index models

Kangning Wang*

Key Laboratory of Group & Graph Theories and Applications, Chongqing University of Arts and Sciences, Chongqing, China
Shandong University Qilu Securities Institute for Financial Studies, Shandong University, Jinan, China

ARTICLE INFO

Article history:
Received 8 September 2014
Accepted 2 April 2015
Available online 28 April 2015

AMS subject classification:
primary 62G05
secondary 62E20

Keywords:
Quantile regression
Variable selection
Oracle property
Single-index structure
High dimension

ABSTRACT

Direction estimation and variable selection in a general class of models with single-index structure are considered. Under mild condition, we show simple linear quantile regression can offer a consistent and asymptotical normal estimate for the direction of index parameter vector in the presence of diverging number of predictors, and it does not need to estimate the link function, and without error distribution constraint. To do variable selection, we penalize the simple linear quantile regression by SCAD, and the oracle property is established. Simulation results and real data analysis confirm our method.

1. Introduction

Let $y \in \mathbb{R}$ be a response variable and $\boldsymbol{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ be a p-dimensional predictor vector, where "T" is transposition. In practice, p is large, and can be assumed to depend on the sample size n at some rate (Fan & Li, 2001; Zhu & Zhu, 2009). Variable selection is a fundamental task for statistical modeling in high-dimensional settings. Traditional variable selection procedures follow either best subset selection or its stepwise variants. However, subset selection is computationally prohibitive when the number of predictors is large. Moreover, as analyzed by Breiman (1996), subset selection may suffer from instability because of its inherent discreteness. To deal with these drawbacks, various penalized methods have been proposed during the past years to perform variable selection and shrinkage estimation simultaneously. In particular, the LASSO Tibshirani (1996), the adaptive-LASSO (ALASSO, Zou, 2006) and the SCAD Fan and Li (2001) are very popular methods with promising computational and statistical properties.

In many applications the linear relationship does not hold. So the use of linear regression to describe the relations between y and $\boldsymbol{x}$ is not suitable. To overcome the risk of model misspecification and the "curse of dimensionality", many models have been proposed. Popular choices include the response transformation model: $g_1(y) = \boldsymbol{x}^T\boldsymbol{\theta}^0 + \epsilon$, and the classical single-index model: $y = g_2(\boldsymbol{x}^T\boldsymbol{\theta}^0) + \epsilon$, or the heteroscedasticity case $y = g_2(\boldsymbol{x}^T\boldsymbol{\theta}^0) + g_3(\boldsymbol{x}^T\boldsymbol{\theta}^0) \times \epsilon$. Here $g_1(\cdot)$ is an unknown monotone function, $g_2(\cdot)$ and $g_3(\cdot)$ are unknown link function, random error ϵ is assumed to be independent of $\boldsymbol{x}$ and $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_p^0)^T$ is the index parameter vector. For more details, see Horowitz (1996) and Wang, Xu, and Zhu (2012). A common feature in the model structure about the above mentioned models is the information of the response can be captured through a single linear combination of the covariates called single-index structure.

* Correspondence to: Key Laboratory of Group & Graph Theories and Applications, Chongqing University of Arts and Sciences, Chongqing, China.
E-mail address: wkn1986@126.com.

http://dx.doi.org/10.1016/j.jkss.2015.04.001
1226-3192/

Thus, in this paper, we consider the following class of models with single-index structure

$$y = G\left(\boldsymbol{x}^T\boldsymbol{\theta}^0, \epsilon\right), \tag{1.1}$$

where $G(\cdot)$ is unspecified link function. Model (1.1) was originally proposed in Li and Duan (1989) and Li (1991), it is equivalent to say that the response y is independent of $\boldsymbol{x}$ given the index $\boldsymbol{x}^T\boldsymbol{\theta}^0$. Clearly, (1.1) is very general and covers the linear models, the transformation linear model, the classical single-index model and its heteroscedastic case mentioned above as a special case. 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 $\boldsymbol{x}^T\boldsymbol{\theta}^0$ can be absorbed into the link function. With a known direction of $\boldsymbol{\theta}^0$, the scatter plot of y versus $\boldsymbol{x}^T\boldsymbol{\theta}^0$ suffices to provide information about $G(\cdot)$ Cook (1998). Thus we are only concerned with the direction of $\boldsymbol{\theta}^0$, regardless of the link function $G(\cdot)$, our aim is to provide a consistent estimator of the direction of $\boldsymbol{\theta}^0$, and it suffices to produce a sparse estimate of the direction of $\boldsymbol{\theta}^0$ for variable selection.

Much research has been done for the family of general single-index models (1.1). For estimating the index $\boldsymbol{\theta}^0$, which include but are not limited to least squares (LS) method (Li & Duan, 1989; Zhu, Qian, & Lin, 2011), structural adaptation method (Dalalyan, Juditsky, & Spokoiny, 2008; Hristache, Juditsky, & Spokoiny, 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, Wang, Zhu, and Ferré (2010) and Zhu et al. (2011). On the other hand, some attempts have also been made to address the variable selection problem for (1.1). Wu and Li (2011) investigated the asymptotic properties of sufficient dimension reduction estimators equipped with SCAD penalty (Fan & Li, 2001), when $p \to \infty$. Zhu and Zhu (2009), Zhu et al. (2011) and Wang, Xu et al. (2012) proposed variable selection methods for (1.1) via penalized LS in the sufficient dimension reduction context. For the classical single index model $y = g_2(\boldsymbol{x}^T\boldsymbol{\theta}^0) + \epsilon$, Kong and Xia (2007) and Naik and Tsai (2001) proposed new selection criteria, Wang and Yin (2008) proposed the sparse MAVE method, Peng and Huang (2011) and Zeng, He, and Zhu (2012) proposed penalized LS methods.

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. Quantile regression Koenker and Gilbert (1978) is a major breakthroughs in the past few decades. It greatly improves the mean regression method in the sense of robustness and comprehensiveness. Many variable selection methods for linear models have been done based on QR, which include but not limited to Belloni and Victor (2011), Zou and Yuan (2008), Wu and Liu (2009) and Wang, Wu, and Li (2012). Also, QR has been used for single index model in the recent literature, see Chaudhuri, Doksum, and Samarov (1997), Wu, Yu, and Yu (2010), Zhu, Huang, and Li (2012), Fan and Zhu (2013), Jiang, Zhou, Qian, and Shao (2012), Jiang, Zhou, Qian, and Chen (2013), Kong and Xia (2012) and Hu, Gramacy, and Lian (2013). But these methods all focus on estimation of the traditional single index model: $y = g_2(\boldsymbol{x}^T\boldsymbol{\theta}^0) + \epsilon$ and the dimension p is fixed, and much less has been done for variable selection. These approaches and asymptotic results, however, cannot be directly extended to the general model (1.1) and "$p \to \infty$" setting.

In this work, index direction estimation and variable selection for models (1.1) in the "$p \to \infty$" setting are considered. We first identify and estimate the direction of $\boldsymbol{\theta}^0$ using quantile regression (QR) (Koenker & Gilbert, 1978), and show the resulting estimate is consistent and asymptotical normal. To select the relevant variables, we then employ the SCAD penalty. This paper makes the following contributions to the literature. (i) Under mild conditions, we show that for any link function $G(\cdot)$ and distribution for the error, with arbitrary $\tau \in (0, 1)$, the τth simple linear QR coefficient for $y|\boldsymbol{x}$ is proportional to $\boldsymbol{\theta}^0$ in model (1.1). (ii) We prove that the resulting SCAD penalized simple linear QR estimates for model (1.1) enjoy the oracle property in the "$p \to \infty$" setting. The theoretical result in (i) implies that we do not need to estimate the involved nonparametric transformation or link function and without distribution constraint for the error, the ordinary linear QR actually results in a consistent estimate for the direction of $\boldsymbol{\theta}^0$. This will bring much convenience for calculation, furthermore, the QR can provide a robust modeling tool for model (1.1). Result in (ii) indicates that the variable selection procedure works very well as if the true relevant variables in model (1.1) were known in advance.

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

2. Methodology and main results

2.1. Direction identification

We first identify the direction of $\boldsymbol{\theta}^0$ in the population level. With regard to model (1.1), we find $y \perp \boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\theta}^0$, where $\perp$ indicates independence, that is, y and $\boldsymbol{x}$ are independent conditioned on $\boldsymbol{x}^T\boldsymbol{\theta}^0$ or equivalently, the conditional distribution of $y|\boldsymbol{x}$ equals that of $y|\boldsymbol{x}^T\boldsymbol{\theta}^0$. In the sufficient dimension reduction setting, Cook (1998) and Zhu et al. (2011) showed that, with a known direction estimator $\bar{\boldsymbol{\theta}}$, the summary plot of y versus $\boldsymbol{x}^T\bar{\boldsymbol{\theta}}$ suffices to provide information about $G(\cdot)$, or say, the plot of y versus $\boldsymbol{x}^T\bar{\boldsymbol{\theta}}$ is an estimated sufficient summary plot, and is often all that is needed to carry on the subsequent analysis. Thus combining with the quantile regression, which is robust and comprehensive, we consider the following population level linear quantile regression loss

$$\mathcal{L}^\tau(b, \boldsymbol{\theta}) := \mathrm{E}\left[\rho_\tau\left(y - b - \boldsymbol{x}^T\boldsymbol{\theta}\right)\right], \tag{2.1}$$