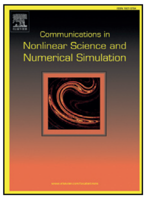
Commun Nonlinear Sci Numer Simulat 36 (2016) 109–128

Contents lists available at ScienceDirect

Commun Nonlinear Sci Numer Simulat

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



Robust check loss-based variable selection of high-dimensional single-index varying-coefficient model

Yunquan Song[a,b,1,*], Lu Lin[b], Ling Jian[a]

[a] College of Science, China University of Petroleum, Qingdao, China
[b] Shandong University Qilu Securities Institute for Financial Studies and School of Mathematics, Shandong University, Jinan, China

ARTICLE INFO

Article history:
Received 18 August 2015
Revised 17 November 2015
Accepted 19 November 2015
Available online 2 December 2015

Keywords:
Basis approximation
Difference Convex algorithm
High-dimensional
Nonlinear
Single-index varying-coefficient model
Variable selection

ABSTRACT

Single-index varying-coefficient model is an important mathematical modeling method to model nonlinear phenomena in science and engineering. In this paper, we develop a variable selection method for high-dimensional single-index varying-coefficient models using a shrinkage idea. The proposed procedure can simultaneously select significant nonparametric components and parametric components. Under defined regularity conditions, with appropriate selection of tuning parameters, the consistency of the variable selection procedure and the oracle property of the estimators are established. Moreover, due to the robustness of the check loss function to outliers in the finite samples, our proposed variable selection method is more robust than the ones based on the least squares criterion. Finally, the method is illustrated with numerical simulations.

1. Introduction

Single-index varying-coefficient model is an important mathematical modeling method of nonlinear phenomena in science and engineering. It has been widely used as a useful generalization of the linear regression to depict inherent behaviors in nonlinear science and complexity. In general, it has the following form

$$Y = \mathbf{g}^T(\boldsymbol{\beta}^T X)Z + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $(X, Z) \in \mathbb{R}^p \times \mathbb{R}^q$ are covariates, Y is the response variable, $\mathbf{g}(\cdot)$ is an q-dimensional vector of unknown functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional unknown parameter vector, and $\boldsymbol{\varepsilon}$ is the model error. For the sake of identifiability, we assume that $||\boldsymbol{\beta}|| = 1$, the first component of $\boldsymbol{\beta}$ is positive, and $\mathbf{g}(x)$ cannot be the form as $\mathbf{g}(x) = \boldsymbol{\alpha}^T x \boldsymbol{\beta}^T x + \boldsymbol{\gamma}^T x + c$, where $||\cdot||$ denotes the Euclidean metric, $\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \mathbb{R}^p, c \in \mathbb{R}$ are constants, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not parallel to each other (Feng and Xue, [7]; Xue and Pang, [28]). The major advantage of model (1.1) is that it does not suffer from the curse of dimensionality which is often encountered in multivariate nonparametric settings, since $\mathbf{g}(\cdot)$ is a vector-valued function of univariate variable. In addition, model (1.1) is flexible enough to cover many important models. For example, if $q = 1$ and $Z = 1$, model (1.1) reduces to the standard single-index model (see Härdle et al., [8]; Stude and Zhu, [20]; Xue and Pang, [28]). If $p = 1$ and $\beta = 1$, model (1.1) is the varying coefficient model (see Breiman, [2]; Fan and Zhang, [6]; Hastie and Tibshirani, [9]; Xue and Zhu, [29]). Thus, model (1.1) is easily

* Corresponding author at: College of Science, China University of Petroleum, Qingdao, China. Tel.: +86 13863900453.
E-mail address: syqfly1980@163.com (Y. Song).

[1] This research was supported by NNSF project (2014010284) of China, NSF project (ZR2014AQ017) of Shandong Province of China, the Fundamental Research Funds for the Central Universities(15CX02083A).

http://dx.doi.org/10.1016/j.cnsns.2015.11.013
1007-5704/© 2015 Elsevier B.V. All rights reserved.

interpreted in real applications because it has the features of both the single-index model and the varying-coefficient model. Because of their flexibility and interpretability, much work has been done on parameter estimation and hypothesis, but mostly in the mean regression setting. To our knowledge, there still lacks of study in the quantile regression setting.

Quantiles themselves can be defined without moment conditions. Quantile regression, which was introduced by Koenker and Basset [12], extends the regression model to conditional quantiles of the response variable. Quantile regression is particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile. Koenker and Basset [12] gave an overview of its applications in statistics and showed its advantages over the least squares regression. First, while the least squares regression can be inefficient if the errors are highly non-normal, quantile regression (QR) is more robust to non-normal errors and outliers. Second, QR also provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of Y, not merely its conditional mean. Third, QR is invariant to monotonic transformations, such as $\log(\cdot)$, a monotone transform of Y, is $h(Q_q(Y))$, and the inverse transformation may be used to translate the result back to Y. This is not possible for the mean as $E[h(Y)] \neq h[E(Y)]$. In spite of remarkable progression in estimation and hypotheses testing for single-index varying-coefficient model (see Wu et al., [24]; Xia and Li, [25]; Xue and Wang, [27]; Xue and Pang, [28]), it is not well understood how to conduct variable selection efficiently for the single-index varying-coefficient model.

Variable selection is important for any regression problem in that ignoring important predictors brings out seriously biased results, whereas including spurious predictors leads to substantial loss in estimation efficiency. Due to their computational efficiency, various shrinkage methods such as the nonnegative garrotte, the least Absolute Shrinkage and Selection Operator (LASSO) and the Smoothly Clipped Absolute Deviation (SCAD) have been used in parametric models and recognized as promising methods to allows us to do estimation and variable selection simultaneously. Furthermore, the past decade has been observed their extensions to semi-parametric and nonparametric models using basis approximation technique. Regarding variable selection in the single-index varying-coefficient model, it is challenging since it has such a complicated multivariate nonlinear structure concerning the nonparametric function vector $\boldsymbol{g}(\cdot)$ and the unknown parameter vector $\boldsymbol{\beta}$. Feng and Xue [7] proposed penalization methods based on the basis function approximations and SCAD penalty for the single-index varying-coefficient model in the mean regression setting. Their proposed method can select significant variables in the parametric components and the nonparametric components simultaneously. It is important to note that their variable selection approach is based on the least squares method, and thus their method inherits all its drawbacks. For example, their method requires a moment assumption. What's more, it is sensitive to outliers in the finite samples and, consequently, it is not robust to outliers in the dependent variable because of the use of least-squares criterion. Therefore, in the presence of outliers, it is desirable to replace the least squares criterion with a robust one. However, to the best of our knowledge, the robust variable selection method for the single-index varying-coefficient model has not been proposed.

Because of the well known robustness properties of the conditional quantile and the fact that conditional quantiles characterize the entire distribution it is of particular interest to develop methods for variable selection in single-index varying-coefficient models. Surprisingly, in quantile regression this problem has found much less attention. In this paper, we propose a variable selection method for estimating coefficient function using a nonparametric approach. We approximate each varying coefficient function with a B-spline basis (Boor, [3]) and consider a penalized check loss function based on the Euclidean norm of the corresponding coefficient vector. Our variable selection method uses a penalization of the norm of each coefficient vector. Therefore, our work shares the same motivation as in Feng and Xue [7]. However, because our interest lies in the estimation of the conditional quantile of the response, we need to use the check loss instead of the squared loss. Therefore, we take a different approach both theoretically and computationally. In particular, the non-differentiability of the check loss function requires us to develop a different computational algorithm. Furthermore, to show its asymptotic properties, we have to adopt a different approach from the one used in mean regression to handle the issues in the proof of our main results, which are given in details in the appendix.

The rest of the article is organized as follows. In Section 2, we first propose the regularized estimation procedure using basis expansion and the SCAD penalty function. Then, we present some theoretical properties of our variable selection procedure, including the consistency of the variable selection and the oracle property of the regularized estimators. In addition, we present a computational algorithm for obtaining the estimator and selection methods for the tuning parameters. In Section 3, we present numerical examples to demonstrate the utilities of the variable selection method via simulations and a real data example. We conclude the paper with Section 4. All the technical proofs are provided in the Appendix.

2. Methodology and main results

Consider the single-index varying-coefficient model in quantile regression

$$Y = \boldsymbol{g}^T(\boldsymbol{\beta}^T X)Z + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $(X, Z) \in \mathbb{R}^p \times \mathbb{R}^q$ are covariates, Y is the response variable, $\boldsymbol{g}(\cdot)$ is an q-dimensional vector of unknown functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a p-dimensional unknown parameter vector, and the model error $\boldsymbol{\varepsilon}$ satisfy $P(\boldsymbol{\varepsilon} \leq 0) = \tau$ for some known constant $\tau \in (0, 1)$. Under this model, $\boldsymbol{g}^T(\boldsymbol{\beta}^T X)Z$ is the conditional τth-quantile of Y given X and Z. We impose no conditions on the heaviness of the tail probability or homoscedasticity of $\boldsymbol{\varepsilon}$.

For model (2.1), to enhance the model fitting accuracy and interpretability, the true regression coefficient vector $\boldsymbol{\beta}^*$ is commonly imposed to be sparse with only a small proportion of nonzeros (Fan and Li, [4]; Tibshirani, [22]). In this section, we propose a variable selection procedure for the single-index varying-coefficient model in quantile regression (2.1) based on the basis