Journal of Computational and Applied Mathematics 308 (2016) 330–345

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Robust exponential squared loss-based variable selection for high-dimensional single-index varying-coefficient model

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

[a] College of Science, China University of Petroleum, Qingdao, China

[b] School of Mathematics, Shandong University, Jinan, China

ARTICLE INFO

Article history:
Received 24 December 2015

MSC:
62G08
62H99

Keywords:
Exponential squared loss
High-dimensional
Robust
Single-index varying-coefficient model
Variable selection

ABSTRACT

Robust variable selection procedure through penalized regression has been gaining increased attention in the literature. They can be used to perform variable selection and are expected to yield robust estimates. In this article, we propose a robust variable selection procedure for high-dimensional single-index varying-coefficient model using penalized exponential squared loss. The proposed procedure simultaneously selects significant covariates with functional coefficients and local significant variables with parametric coefficients. With proper choices of penalty functions and regularization parameters, we show the asymptotic normality of the resulting estimate and further demonstrate that the proposed procedures perform as well as an oracle procedure. Our simulation studies reveal that our proposed method performs similarly to the oracle method in terms of the model error and the positive selection rate even in the presence of influential points. In contrast, other existing procedures have a much lower noncausal selection rate. Our analysis unravels the discrepancies of using our robust method versus the other penalized regression method, underscoring the importance of developing and applying robust penalized regression methods.

1. Introduction

Semiparametric regression is a kind of important mathematical modeling method of nonlinear phenomena in science and engineering. With the development of applied sciences, semiparametric regression models have been well researched and popularly used for their flexibility and interpretability. Among semiparametric models, single-index varying-coefficient model is one class of commonly-used models because they effectively avoid the "curse of dimensionality" of nonparametric model and have the explanatory power of the parametric model. In general, it has the following form

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

where $(X, Z) \in \mathbb{R}^p \times \mathbb{R}^q$ are covariates, Y is the response variable, $\boldsymbol{g}(\cdot)$ is a 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 $\boldsymbol{\varepsilon}$ satisfies $E(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2$. For the sake of identifiability, we assume that $\|\boldsymbol{\beta}\| = 1$, the first component of $\boldsymbol{\beta}$ is positive, and $\boldsymbol{g}(x)$ cannot be the form as $\boldsymbol{g}(x) = \boldsymbol{\alpha}^T x\boldsymbol{\beta}^T x + \boldsymbol{\gamma}^T x + c$, 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 (see [1,2]). It is important to emphasize that model (1.1) is flexible enough to cover many important

* Corresponding author at: College of Science, China University of Petroleum, Qingdao, China.
E-mail address: syqfly1980@upc.edu.cn (Y. Song).

http://dx.doi.org/10.1016/j.cam.2016.05.030
0377-0427/

models such as the standard single-index model (see [3–7]) and the varying coefficient model (see [8–12]). Thus, model (1.1) is easily interpreted in real applications because it has the features of both the single-index model and the varying-coefficient model.

Although much work has been done on parameter estimation and hypothesis, it is not well understood how to conduct variable selection efficiently for the single-index varying-coefficient model. We all know that single-index varying-coefficient model is one class model of semiparametric model. Variable selection for semiparametric regression models consists of two components: model selection for nonparametric components and selection of significant variables for the parametric portion. Thus, semiparametric variable selection is much more challenging than parametric variable selection (e.g., linear and generalized linear models) because traditional variable selection procedures including stepwise regression and the best subset selection now require separate model selection for the nonparametric components for each submodel. This leads to a very heavy computational burden. Regarding variable selection in the single-index varying-coefficient model, Feng and Xue [1] proposed penalization methods based on the basis function approximations and SCAD penalty for the single-index varying-coefficient model. 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. It is well known that the least squares method 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. Robust variable selection procedure through penalized regression has been gaining increased attention in the literature. They can be used to perform variable selection and are expected to yield robust estimates. However, to the best of our knowledge, the robust variable selection method for the single-index varying-coefficient model has not been proposed.

In the regression setting, the choice of loss function determines the robustness of the resulting estimators. Thus, for model (1.1), the loss function $\rho(\cdot)$ is critical to the robustness, and for this reason, a loss function with superior robustness performance is of great interest. We know that the exponential squared loss is robust which has the following form

$$\psi_\eta(t) = 1 - \exp(-t^2/\eta),$$

where η is a tuning parameter. The tuning parameter η controls the degree of robustness for the estimators. When η is large, $\psi_\eta(t) \approx t^2/\eta$, and therefore the proposed estimators are similar to the least squares estimators in the extreme case. For a small η, observations with absolute values of $t_i = Y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ will result in large losses of $\psi_\eta(t_i)$ and therefore have a small impact on the estimation of $\boldsymbol{\beta}$. Hence, a smaller η would limit the influence of an outlier on the estimators, although it could also reduce the sensitivity of the estimators. This exponential loss function has been used in AdaBoost for classification problem (see [13]) and in variable selection for the linear regression model (see [14]).

In this article, we discuss how to select η so that the corresponding penalized regression estimators are robust and possess desirable finite and large sample properties. We show that our estimators satisfy selection consistency and asymptotic normality. Through the theoretical and simulation results, we demonstrate the merits of our proposed method.

The rest of this article is organized as follows. In Section 2, we introduce the penalized robust regression estimators with the exponential squared loss based on the basis function approximations, and investigate the sampling properties. In Section 3, we study the robustness properties of our proposed method. In Section 4, algorithm and the choice of tuning parameters are introduced. In Section 5, numerical simulations are conducted to compare the performance of the proposed method with least squares loss and exponential squared loss using the oracle method as the benchmark. In Section 6, we apply the single-index varying-coefficient model and the corresponding estimation method to one real dataset. The proofs are given in the Appendix.

2. Methodology

In this section, we propose a variable selection procedure for the single-index varying-coefficient model (1.1) based on the basis function approximation and SCAD penalty function. First, we use the B-spline functions to approximate the unknown coefficient functions in the model. Then combining with the restraint $\|\boldsymbol{\beta}\| = 1$, we adopt the 'delete-one-component' method proposed by Yu and Ruppert [15] and Feng and Xue [1] to construct the penalized exponential squared loss function.

2.1. Basis function expansion

Suppose that $\{(X_i, Z_i, Y_i),\ 1 \le i \le n\}$ is a sample from (1.1), i.e.,

$$Y_i = \boldsymbol{g}^T(\boldsymbol{\beta}^T X_i)Z_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $X_i = (X_{i1}, \ldots, X_{ip})^T$ and $Z_i = (Z_{i1}, \ldots, Z_{iq})^T$, and ε_i's are unobservable random errors with mean 0 and finite variance σ^2. Assume that $\{\varepsilon_i,\ 1 \le i \le n\}$ are independent of $\{(X_i, Z_i),\ 1 \le i \le n\}$.

Since $\boldsymbol{g}(\cdot)$ is unknown, similar to He et al. [16] and Feng and Xue [1], we replace $\boldsymbol{g}(\cdot)$ by its basis function approximations. More specifically, let

$$B(u) = (B_1(u), \ldots, B_L(u))^T$$