Journal of the Korean Statistical Society 44 (2015) 45–57

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Robust variable selection in partially varying coefficient single-index model

Huiming Zhu[a], Zhike Lv[a,*], Keming Yu[b], Chao Deng[a]

[a] College of Business Administration, Hunan University, Changsha, 410082, PR China

[b] Department of Mathematical Sciences, Brunel University, London UB8 3PH, UK

ARTICLE INFO

Article history:
Received 15 October 2013
Accepted 7 May 2014
Available online 2 June 2014

AMS 2000 subject classifications:
primary 62G08
secondary 62G35

Keywords:
Variable selection
Spline approximation
Modal regression
SCAD
Oracle property

ABSTRACT

By combining basis function approximations and smoothly clipped absolute deviation (SCAD) penalty, this paper proposes a robust variable selection procedure for a partially varying coefficient single-index model based on modal regression. The proposed procedure simultaneously selects significant variables in the parametric components and the nonparametric components. With appropriate selection of the tuning parameters, we establish the theoretical properties of our procedure, including consistency in variable selection and the oracle property in estimation. Furthermore, we also discuss the bandwidth selection and propose a modified expectation–maximization (EM)-type algorithm for the proposed estimation procedure. The finite sample properties of the proposed estimators are illustrated by some simulation examples.

1. Introduction

Partially varying coefficient single-index models (PVCSIMs) combine naturally the advantages of both the single-index models and the varying coefficient models. Ever since Wong, Ip, and Zhang (2008) proposed the PVCSIM, studies in this class of models have raised great interest of research in Statistics field. We formulate a PVCSIM as

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

where Y is a response variable, X and Z are of dimensions $p \times 1$ vectors and $q \times 1$ vectors, $\theta(\cdot) = (\theta_1(\cdot), \ldots, \theta_q(\cdot))^T$ is a vector of unknown function, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown parameters, $g(\cdot)$ is an unknown link function, and ε is random error with mean zero. Due to the curse of dimensionality, we assume, for simplicity, that U is univariate. And we also assume that $\|\beta\| = 1$ and $\mathrm{sigh}(\beta_1) = 1$ to ensure identifiability, where $\|\cdot\|$ denotes the Euclidean metric.

Model (1.1) is quite flexible enough to cover a variety of existing statistical models. For example, if $g(\cdot) = 0$, it reduces to the standard varying-coefficient model. When $\theta(\cdot)$ is an unknown constant parameter, then model (1.1) is a partially linear single-index model (Wang & Wu, 2013; Wang, Xue, Zhu, & Chong, 2010). In addition, model (1.1) becomes the standard single-index model when $Z = 0$ or $\theta(\cdot) = 0$ (Wu, Yu, & Yu, 2010). Due to its flexibility and generality, model (1.1) has gained much attention in recent years. Wang and Xue (2011) developed a stepwise approach to obtain asymptotic normality estimators of the varying-coefficient vector and the parametric vector. Huang and Zhang (2010) constructed the

[*] Corresponding author. Tel.: +86 731 88684461.
E-mail address: lzk2011@hnu.edu.cn (Z. Lv).

http://dx.doi.org/10.1016/j.jkss.2014.05.002

confidence region for the parameter β in model (1.1) based on the empirical likelihood technique. While Huang (2011) used the empirical likelihood method to study the confidence regions of the varying-coefficient parts. Huang, Lin, Feng, and Pang (2013) proposed a class of efficient penalized estimating equations to estimate the index parametric components in the PVCSIM. Feng and Xue (2013) also considered the problem of variable selection in the PVCSIM. However, the aforementioned existing researches were mainly built on either the least-square or empirical likelihood method, which are expected to be very sensitive to the outliers and its efficiency may be significantly decreased for many commonly used non-normal errors.

Recently, Yao, Lindsay, and Li (2012) proposed a new estimation approach based on a local modal regression for the nonparametric model. Then, Zhang, Zhao, and Liu (2013) and Zhao, Zhang, Liu, and Lv (2014) investigated the partially linear varying coefficient model based on modal regression, respectively. And Liu, Zhang, Zhao, and Lv (2013) developed a new robust and efficient estimation procedure based on local modal regression for single index models. A distinguishing characteristic of their method is that it introduces an additional tuning parameter which is automatically selected using the observed data to achieve both robustness and efficiency of the resulting estimate. Namely, their method is not only robust when there are outliers or the error distribution is heavy-tail, but as asymptotically efficient as the ordinary least-square-based estimator when the data include no outliers and the error distribution is a Gaussian distribution. Due to its nice property, it has attracted increasing attention. Here, we extend the modal regression approach to the model (1.1).

Variable selection is a crucial issue in regression analysis. In practice, a number of variables are available for inclusion in an initial analysis, but many of them may not be significant and should be excluded from the final model to increase the accuracy of prediction. Traditional variable selection methods such as stepwise regression and best subset selection are computationally infeasible when the number of predictors is large. Therefore, various shrinkage methods such as the LASSO (Tibshirani, 1996), the adaptive LASSO (Zou, 2006) and the SCAD (Fan & Li, 2001) have gained much attention in recent years. However, the LASSO is known to be near mini-max optimal as well as consistent under certain regularity conditions, Zou (2006) showed that it falls short of attaining the oracle property. By this property, an estimator estimates a zero coefficient exactly as zero with probability approaching one, while still being asymptotically normal for the non-zero coefficients in large samples. In this respect, the LASSO is inferior to the SCAD estimator which possesses the oracle property. So in the present paper, we prefer the SCAD of Fan and Li (2001) since it simultaneously satisfies the mathematical conditions for unbiasedness, sparsity, and continuity. More details can be found in Fan and Li (2001). Since the SCAD was proposed, there has been a large number of literature focused on its applications in many important nonparametric and semiparametric models.

In this paper, we investigated the variable selection for the varying coefficient function $\theta(\cdot)$ and the unknown parametric index β in model (1.1) based on modal regression. By combining the basis function approximate and the SCAD penalty, we develop a variable selection procedure for PVCSIM. More specifically, we first use the B-spline functions to approximate the unknown coefficient functions and link function in model (1.1). And then combine with the restraint $\|\beta\| = 1$ to construct the penalized estimation function for PVCSIM based on modal regression. Under certain regularity conditions, we are able to establish this variable selection procedure is consistent, and the estimators have oracle property. Moreover, a modified version of a modal expectation–maximization (MEM) algorithm is proposed to obtain the solutions for the object function. Some simulation studies show that, when data is contaminated by outliers, the proposed variable selection procedure can perform well in finite samples.

The layout of the remainder of the paper is as follows. In Section 2, following the idea of the modal regression approach, we propose the regularized estimation produce using basis expansion and the SCAD penalty function. Then, under some regularity conditions, we establish some theoretical properties of the proposed variable selection procedure. We describe the details of bandwidth selection and propose a modified MEM algorithm. In addition, we give the method of choosing the tuning parameters in Section 3. In Section 4, we conduct some simulation studies to examine the finite sample performance of the proposed procedures. Finally, in Section 5 we conclude the paper. All the regularity conditions and the technical proofs are relegated to the Appendix.

2. Estimation and variable selection procedure

As a measure of center, the median and mode have the common advantage of robustness, when there exist outliers. Furthermore, since the modal regression focuses on the relationship between the majority data points and summaries the “most likely”conditional values, it can provide more meaningful point prediction than the mean regression when the error density is skewed. Suppose that $\{(Y_i, X_i, Z_i, U_i), i = 1, \ldots, n\}$ is an i.i.d. sample from model (1.1). Then following the method of Yao et al. (2012), the robust modal estimate of the PVCSIM is to maximize

$$\frac{1}{n}\sum_{i=1}^{n}\phi_h\Big(Y_i - Z_i^T\theta(U_i) - g(X_i^T\beta)\Big), \tag{2.1}$$

where $\phi_h(t) = h^{-1}\phi(t/h)$, $\phi(t)$ is a kernel density function, and the choice of $\phi(\cdot)$ is not very important. h is a bandwidth. For ease of computation, we use the standard normal density for $\phi(t)$ throughout the present article.

Remark 1. The choice of kernel is not very important because it is possible to obtain estimators with somewhat improved asymptotic properties by using different kernels (see, e.g., Eddy, 1980; Romano, 1988). For the simplicity of the calculation, we use the Gaussian density for $\phi(t)$.