Journal of Statistical Planning and Inference 176 (2016) 1–21

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Composite quantile regression and variable selection in single-index coefficient model[☆]

Riquan Zhang[a,b,*], Yazhao Lv[a], Weihua Zhao[c], Jicai Liu[d]

[a] School of Finance and Statistics, East China Normal University, Shanghai, 200241, PR China
[b] Department of Mathematics, Shanxi Datong University, Datong, 037009, PR China
[c] School of Science, Nantong University, Nantong, 226019, PR China
[d] College of Mathematics and Sciences, Shanghai Normal University, Shanghai, 200234, PR China

ARTICLE INFO

Article history:
Received 23 March 2015
Received in revised form 14 April 2016
Accepted 15 April 2016
Available online 25 April 2016

MSC:
primary 62G05
secondary 62E20

Keywords:
Single index coefficient model
Composite quantile regression
Asymptotic normality
Asymptotic relative efficiency
Variable selection
Adaptive LASSO

ABSTRACT

In this paper, we propose a composite minimizing average check loss estimation procedure for composite quantile regression (CQR) in the single-index coefficient model (SICM). The asymptotic normalities of the proposed estimators are established, and the asymptotic relative efficiencies (ARE) of the proposed estimators compared with those by least square method are also discussed. We further investigate a variable selection procedure by combining the proposed estimation method with adaptive LASSO penalized method in CQR of SICM. The oracle property of the proposed variable selection method is also established. Simulations with various non-normal errors and one real data application are conducted to assess the finite sample performance of the proposed estimation and variable selection methods.

1. Introduction

Consider the following single-index coefficient model (SICM)

$$Y = \alpha_0(\mathbf{X}^T\theta) + \alpha(\mathbf{X}^T\theta)^T\mathbf{Z} + \varepsilon, \tag{1}$$

where Y is the response variable, $\mathbf{X} = (X_1, \ldots, X_p)^T \in R^p$, $\mathbf{Z} = (Z_1, \ldots, Z_d) \in R^d$ are the covariates, $\alpha_0(\cdot)$ is baseline scalar function, $\alpha(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_d(\cdot))^T$ are d-dimensional coefficient functions, θ is unknown index parameter and ε is the model error. For sake of identification, we set $\|\theta\| = 1$ and the first elements of θ is positive.

The SICM (1) proposed by Xia et al. (1999), is very general and includes many common models as its special case. When $\mathbf{X}$ is scalar variable, model (1) is the varying coefficient model (VCM) proposed by Hastie and Tibshirani (1993). When $\mathbf{Z} \equiv 1$, model (1) reduces to the single-index model.

[☆] The research was supported in part by National Natural Science Foundation of China (11571112, 11501372), National Social Science Fund of China (15BTJ027), Doctoral Fund of Ministry of Education of China (20130076110004), Program of Shanghai Subject Chief Scientist (14XD1401600), 111 Project of China (B14019), Shanghai Municipal Science and Technology Research Project (14DZ1201902) and Natural Science Fund of Nantong University (14B28).

[*] Corresponding author at: School of Finance and Statistics, East China Normal University, Shanghai, 200241, PR China.
E-mail address: zhangriquan@163.com (R. Zhang).

http://dx.doi.org/10.1016/j.jspi.2016.04.003
0378-3758/© 2016 Elsevier B.V. All rights reserved.

Xia et al. (1999) extended the method of Härdle et al. (1993) in single-index model to SICM and obtained the estimation of θ by minimizing simultaneously objective function with respect to θ and h. They showed that the estimator of θ achieves the best convergence rate and avoids the "undersmoothing" condition. However, they estimated the index parameter θ by a computationally expensive least-squares cross-validation methods, which is not practice in reality.

Fan et al. (2003) considered a special example of SICM and proposed an efficient estimation method with bandwidth selection strategy. Lu et al. (2007) established the asymptotic theory of the profile likelihood estimation in SICM. Recently, Xue and Pang (2012) proposed an estimation method by estimation equation and obtained the confidence region of the nonparametric coefficient function. Huang and Zhang (2012a) derived a confidence interval of the index parameter θ in SICM by profile empirical likelihood method. Huang and Zhang (2012b) considered the testing of θ in SICM by applying the generalized likelihood ratio test proposed by Fan et al. (2001). More recently, Feng and Xue (2013) proposed an estimation of SICM by spline method and further discussed the variable selection of the parameter and the nonparametric coefficient functions.

All the estimation methods mentioned above focused on the mean regression in SIM, based on least square method or likelihood approach. Since mean regression may loss efficiency when the error distribution is non-normal, the quantile regression proposed by Koenker and Basset (1978) can be viewed as an alternative approach to explore the underlying relationship of the response and the multidimensional covariates. There exist some papers in literature of the quantile regression of the simplified form of SICM, such as the SIM (see Wu et al., 2010; Jiang et al., 2012) and VCM (see Honda, 2004; Kim, 2007; Cai and Xu, 2008). Recently, the quantile regression and variable selection of SICM have been considered by Lv et al. (2014).

Since the estimation efficiency may fluctuate with the particular value of the quantile, Zou and Yuan (2008) proposed composite quantile regression(CQR) by combining the information from multiple quantile regression to obtain more efficient estimator. Composite quantile regression has been proved to be a powerful and robust alternative approach in linear models (Zou and Yuan, 2008) and semi-parametric varying coefficient partially linear models (Kai et al., 2011), when the error distribution deviates far from normal distribution. Though there have been much works on the CQR of various models, no research has been done on the CQR in SICM.

The motivation of this article also comes from the analysis of the environmental dataset from New Territories East in Hong Kong between January 1, 2000 and December 31, 2000. There are six environmental pollutants including X_1: sulfur dioxide concentration (in g/m^3), X_2: respirable suspended particulate concentration (in g/m^3), X_3: nitrogen oxide concentration (in g/m^3), X_4: nitrogen dioxide concentration (in g/m^3), X_5 ozone concentration (in g/m^3) and X_6: water turbidity (in mg/m^3) together with two environment factors—temperature (in Celsius) Z_1 and relative humidity (%) Z_2. In this real data, it is of interest to explore how the pollutant concentrations and environment factors influence the number of daily total hospital admissions (Y) for respiratory diseases. In view of measurements for multiple pollutants, it is useful to construct a single pollutant index to be used in predicting hospital admissions and thus a single-index model is appropriate. To further study the interaction effect between pollutants and two environment factors for the response variable, we analyze this data by SICM with robust CQR method.

Since in real data analysis, irrelevant variables may often be included in the covariates, variable selection has been a fundamental part in multivariate statistical modeling. Feng and Xue (2013) have considered variable selection of SICM in the mean regression. For the real data mentioned above, we also want to determine whether each pollutant has important effect on the number of daily respiratory diseases. We will apply our proposed variable approach to the environmental dataset in the framework of CQR method.

To achieve this goal, in this paper, we will propose a composite minimizing average check loss estimation (CMACLE) procedure to conduct CQR in SICM. We will show that the estimator of θ achieves the best convergence rate without the "undersmoothing" condition. The asymptotic normalities of the proposed estimators are established. In addition, we compare asymptotic relative efficiency of the CMACLE with the mean regression by profile-likelihood method (Lu et al., 2007). Furthermore, we proposed an adaptive LASSO penalized CMACLE procedure to conduct the variable selection the CQR in SICM. The corresponding oracle properties are also established. To demonstrate the theoretical property of the proposed methods, we consider two simulations with various distributed errors. The simulation results and real data analysis are further used to illustrate the performance of our newly proposed method.

The rest of the paper is organized as follows. In Section 2, we outline the estimation procedure and present the algorithm to conduct the composite quantile regression in SICM. Asymptotic results for the proposed estimators and ARE are presented in Section 3. In Section 4, we propose the variable selection procedure and establish the corresponding oracle property. In Section 5, Monte Carlo simulations with various distributed errors and Hong Kong environmental data analysis are conducted to assess the merits of our methods. All the regular conditions and technical proofs are relegated to the Appendix.

2. Estimation methodology

Let $\rho_\tau(u) = u[\tau - I(u < 0)]$ be the check loss function for $\tau \in (0,\ 1)$. Quantile regressions are usually used to estimate the conditional quantile of the response variable Y, which is defined as

$$q_\tau(\mathbf{x}, \mathbf{z}) = \text{argmin}_a \mathrm{E}\left\{\rho_\tau(Y - a)|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}\right\}.$$