Computational Statistics and Data Analysis 70 (2014) 241–256

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Variable selection and semiparametric efficient estimation for the heteroscedastic partially linear single-index model

Peng Lai [a], Qihua Wang [b,c,*], Xiao-Hua Zhou [d]

[a] School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China
[b] Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China
[c] Institute of Statistical Science, Shenzhen University, Shenzhen 518060, China
[d] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

ARTICLE INFO

Article history:
Received 30 January 2012
Received in revised form 7 September 2013
Accepted 16 September 2013
Available online 23 September 2013

Keywords:
Heteroscedasticity
Variable selection
Estimating equations
Efficient score function
Oracle property

ABSTRACT

An efficient estimating equations procedure is developed for performing variable selection and defining semiparametric efficient estimates simultaneously for the heteroscedastic partially linear single-index model. The estimating equations are proposed based on the smooth threshold estimating equations by using the efficient score function of partially linear single-index models. And this estimating equations procedure can be used to perform variable selection without solving any convex optimization problems, and automatically eliminate nonsignificant variables by setting their coefficients as zero. The resulting estimators enjoy the oracle property and are semiparametrically efficient. The finite sample properties of the proposed estimators are illustrated by some simulation examples, as well as a real data application.

1. Introduction

Understanding characteristics of subgroups at greatest risk of progression to Alzheimer' Disease (AD) will streamline the drug development process by allowing us to target interventions to high risk subjects. How to identify those high-risk subjects is difficult due to the mixed process of normal aging and AD disease progression. Imaging technology is playing an increasing role in identifying risk factors for predicting disease progression. Particularly, positron emission tomography (PET) can detect glucose metabolism in the brain using a radioactive compound (PiB) for measuring brain beta-amyloid. However, there are many region-level variables from a PET scan, and some of them are associated with the cognitive decline and some are not. It is a challenge to select those important PET variables, which are associated with cognitive decline of a subject. To answer this question, we use one of the largest imaging datasets from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Regions of Interests (ROIs) computed for PET data include parietal (angular and supramarginal gyrus), temporal (mid-temporal), hippocampus, and lateral prefrontal cortex, result in a total of 12 variables. The Mini-Mental State Exam (MMSE) is also completed. The outcome variable is the change MMSE score from the baseline to the two-year follow-up time. Since some of these variables may be linearly related to the outcome, while the others are not, we consider the following heteroscedastic partially linear single-index model for the data:

$$Y = \theta^{\top} Z + g(X^{\top}\beta) + \varepsilon, \tag{1}$$

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing 100190, China. Tel.: +86 10 62641688; fax: +86 15101512088.
E-mail address: qhwang@amss.ac.cn (Q. Wang).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.09.012

where Y is a scalar response variable, $(X, Z) \in R^p \times R^q$, $g(\cdot)$ is an unknown univariate link function, ε is the random error with $E(\varepsilon|X, Z) = 0$, and $\mathrm{var}(\varepsilon|X, Z) = \sigma^2(X, Z)$, $\sigma^2(\cdot)$ is an unknown function. We can see that the conditional variance is correlated with the covariates X and Z, and heteroscedasticity is very common in practice. (β, θ) is an unknown vector in $R^p \times R^q$ with $\|\beta\| = 1$ (where $\|\cdot\|$ denotes the Euclidean metric). Denote the true parameters as β_0 and θ_0.

This model has gained much attention in recent years. See, e.g., Carroll et al. (1997), Xia et al. (1999), Xia and Härdle (2006) and Zhu and Xue (2006). However, their methods assume that all covariates (X, Z) contain useful information to predict the response variable. If irrelevant covariates are included in the model, which is very common in high-dimensional environments, the values of estimated parameters as well as the accuracy of prediction will suffer. Therefore, it is natural to exclude the irrelevant covariates from the partially linear single-index model. To eliminate the irrelevant variables, variable selection methods have been widely used in recent years. Various methods for various models have been studied, examples include LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006), Dantzig selector (Candes and Tao, 2007), smooth threshold estimating equations (Ueki, 2009) and so on. A single-index model is a special case of model (1) when $\theta = \mathbf{0}$. For the single-index model, various variable selection techniques are developed. Naik and Tsai (2001) considered variable selection using slice inverse regression in which the predictor X is required to be continuous and elliptically symmetric. Kong and Xia (2007) suggested a separated cross validation method to select the variables. Zhu and Zhu (2009) combined the sufficient dimension reduction with non-concave penalty to do variable selection. Model (1) reduces to a partially linear model when $p = 1$ and $\beta = 1$. For a partially linear model with measurement errors, Liang and Li (2009) proposed the penalized least squares and the penalized quartile regression procedures for variable selection, where the nonconvex penalized principle is used. Zhao and Xue (2009) combined basis function approximations with SCAD penalty to construct a variable selection procedure for semiparametric varying coefficient partially linear models.

The standard way to do variable selection is to add a penalty item to a loss function, that is $L(\beta)+\sum_{j=1}^{d} \rho_j(|\beta_j|)$, where $L(\beta)$ represents the loss function and ρ_j represents the nonnegative penalty function for the jth parameter, and d is the dimension of the parameter β of interest. The procedure of variable selection reduces to a convex optimization problem. As we know, in general this procedure is computationally expensive. To avoid the convex optimization problems, Ueki (2009) proposed the smooth threshold estimating equations and applied this method to a simulation study of multiple linear regression with censored responses. This method is easily implemented with the Newton Raphson algorithm, yields sparse solutions without solving any convex optimization problems, and the resulting estimators possesses the oracle property in the sense of Fan and Li (2001). In this paper, we develop an efficient estimating equations procedure for performing variable selection and defining semiparametric efficient estimates simultaneously for the heteroscedastic partially linear single-index model. The estimating equations are developed based on the smooth threshold estimating equations due to Ueki (2009) with the efficient score function of the partially linear single-index model. We found that Liang et al. (2010) also considered the variable selection problem for the partially linear single index model. Liang et al. (2010) used a penalized profile least squares approach to simultaneously estimate parameters and select important variables, where the SCAD penalty was used. This is different from the estimating equation method developed here, which is computed more efficiently. Also, they consider the special case where $(X^\top, Z^\top)^\top$ and ε are independent for the partial linear single index model. This assumption is too strong in reality. This is different from the nonparametric heteroscedastic case considered here.

The rest of this article is organized as follows. In Section 2, we develop a variable selection approach based on the smooth threshold estimating equations, and prove the consistency of the variable selection and asymptotic normality of the resulting estimators. An asymptotically semiparametric efficient problem is also discussed in Section 2. In Section 3, some simulation studies are conducted to investigate the finite sample properties of the proposed estimators and check the consistency of variable selection. We also apply our variable selection procedure to a real data analysis to identify the important variables related with Alzheimer' Disease (AD). The proofs of the asymptotic properties are outlined in the Appendix.

2. Methodology and main results

Ueki (2009) proposed the smooth-threshold estimating equations to do variable selection, which only needs a set of estimating equations without specifying a loss function. This method avoids the convex optimization problems and the derivative of a loss function is not necessary. As Ueki (2009) points out, if we have the $d-$dimensional estimating function $u(\cdot)$ for estimating η_0 with $E[u(X; \eta_0)] = 0$ and $E[\|u(X; \eta_0)\|^2] < \infty$, we can construct the following smooth-threshold estimating equations:

$$(I_d - D)u(\eta) + D\eta = 0, \tag{2}$$

where $u(\eta) = \sum_{i=1}^{n} u(X_i; \eta)$, I_d is the d-dimensional identity matrix, matrix D is a diagonal matrix, and the diagonal elements are $(\hat{\delta}_1, \ldots, \hat{\delta}_d)$ with $\hat{\delta}_i = \min(1, \frac{\lambda}{|\tilde{\eta}_i|^{1+\gamma}})$, $\tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_d)^\top$ being a root-n consistent estimator of η_0, and λ and γ the penalty tuning parameters. Thus, we can solve this estimating equation on η to obtain the estimator of parameter vector η_0. From the definition of $\hat{\delta} = (\hat{\delta}_1, \ldots, \hat{\delta}_d)$, it follows that when $\hat{\delta}_i = 1, i \in \{1, \ldots, d\}$, the solution of (2) for η_i is that $\hat{\eta}_i = 0$. Therefore, estimating Eq. (2) can find the nonsignificant variables among all candidate variables automatically, yielding a sparse solution, completing a variable selection procedure and parameters estimation simultaneously. Eq. (2) can be solved easily with familiar techniques, such as the Newton–Raphson algorithm, and no convex optimization is required.