Journal of Multivariate Analysis 139 (2015) 227–244

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Model detection and estimation for single-index varying coefficient model

Sanying Feng [a,b,*], Liugen Xue [a]

[a] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China
[b] School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China

ARTICLE INFO

Article history:
Received 23 June 2014
Available online 28 March 2015

AMS subject classification:
primary 62G05
secondary 62G20

Keywords:
Single-index varying coefficient model
Combined penalization
Model detection
Oracle property
MAVE

ABSTRACT

Single-index varying coefficient model (SIVCM) is a powerful tool for modeling nonlinearity in multivariate estimation, and has been widely used in the literature due to the fact that it can overcome the well-known phenomenon of "curse-of-dimensionality". In this paper, we consider the problem of model detection and estimation for SIVCM. Based on the proposed combined penalization procedure, we can identify the true model structure consistently, and obtain a new semiparametric model—partially linear single-index varying coefficient model (PLSIVCM). Under the appropriate conditions, we demonstrate that the proposed penalized estimators of parametric and nonparametric components of PLSIVCM are consistent, but their asymptotic distributions are not available. Hence, we extend the minimum average variance estimation method to PLSIVCM, and establish the asymptotic normality for the refined estimators of index parameters, constant coefficients and varying coefficient functions, respectively. The finite sample performances of the proposed methods are illustrated by a Monte Carlo simulation study and the real data analysis.

1. Introduction

Single-index varying coefficient model is a class of important semiparametric model proposed in recent years, and has attracted lots of attention due to its features of both the traditional single-index model (see [8]) and the varying coefficient model (see [9,6]). The single-index varying coefficient model (SIVCM) assumes the following structure:

$$Y = \sum_{j=1}^{q} g_j(\beta^\tau X)Z_j + \varepsilon, \tag{1.1}$$

where Y is the response variable, X and $Z = (Z_1, \ldots, Z_q)^\tau$ are covariates, $\beta = (\beta_1, \ldots, \beta_p)^\tau$ is a $p \times 1$ vector of unknown index parameters, $\mathbf{g}(\cdot) = (g_1(\cdot), \ldots, g_q(\cdot))^\tau$ is a $q \times 1$ vector of unknown coefficient functions and the model error ε has mean 0 and finite variance σ^2. Generally, Z_1 may be taken as 1 so that the model has an intercept function term. For the sake of identifiability, we assume that $\|\beta\| = 1$ and the first component of β is positive, where $\|\cdot\|$ denotes the Euclidean metric. Moreover, when $Z = X$ or $Z = (1, X^\tau)^\tau$, we need the additional identifiability condition, i.e., $G(x, z)$ cannot be the form as below

$$G(x, z) = \beta^\tau x \alpha^\tau x + \gamma^\tau x + c,$$

* Corresponding author.
E-mail address: fsy5801@sina.com (S. Feng).

http://dx.doi.org/10.1016/j.jmva.2015.03.008
0047-259X/

where $G(x, z) = \sum_{j=1}^{q} g_j(\beta^\tau x)z_j$, $\alpha, \gamma \in R^p$, $c \in R$ are constants, and α and β are not parallel to each other.

Model (1.1) has two main advantages, one 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 function vector of univariate variable; the other is that it includes the nonlinear interaction effects between the covariates X and Z. There are various approaches to estimate the regression coefficients in model (1.1). For example, Xia and Li [26] investigated a class of single-index coefficient regression models, which include model (1.1) as a special example. Under the case that $Z = (1, X^\tau)^\tau$, Fan et al. [5] established a new computationally efficient procedure based on the profile least-squares local linear weighted regression. Xue and Wang [29] proposed an empirical likelihood method, and constructed the empirical likelihood confidence regions for the index parameter vector of model (1.1). Xue and Pang [28] obtained estimates of nonparametric and parametric components by using local linear smoothing and estimating equations, and showed the asymptotic normality and optimal uniform convergence rate of the estimators. Feng and Xue [7] presented a variable selection procedure based on the basis function approximations and the smoothly clipped absolute deviation (SCAD, Fan and Li [4]) penalty for model (1.1). Huang et al. [12] proposed a penalized estimating equations method to simultaneously identify significant variables of the index and estimate the nonzero coefficients of the index parameters of model (1.1).

In the above studies, they all assumed that the nonlinear interaction effects between the covariates X and Z are existent. But in practice this is not always true. If the interaction effect between X and Z_j is misspecified, it cannot only increase the complexity of the model but also reduce the estimation accuracy. Since the optimal parametric estimation rate is $n^{-1/2}$ and the optimal nonparametric estimation rate is $n^{-2/5}$, treating a parametric component as a nonparametric component can over-fit the data and lead to efficiency loss. Therefore, structure identification is important to model (1.1). Huang [11] constructed a generalized likelihood ratio (GLR) statistic to test if $g_j(\cdot)$ is an unknown constant. Huang et al. [12] proposed a new generalized testing statistic to test whether $g_j(\cdot)$ is a pre-specified parametric function. However, these methods pose great challenges. On one hand, for each submodel, it is necessary to estimate the coefficient functions and index parameters, and that will dramatically increase the computational burden. On the other hand, the critical values of their test statistics are difficult to obtain. Taking these issues into account, we recommend using the penalization method to detect the true structure and estimate the unknown parametric and nonparametric components for model (1.1). The idea of employing penalization methods to identify the model structure has been already adopted by some authors in the literature. For example, Zhang et al. [32] and Huang et al. [13] proposed two methods for determining the linear and nonlinear components in partially linear models; Leng [15] and Hu and Xia [10] presented penalization methods to identify the constant and varying effects in varying coefficient (VC) models. However, these methods cannot do variable selection. Tang et al. [22] proposed a unified variable selection method for VC models, which can identify the varying and constant coefficients and select the significant variables simultaneously. Wang and Kulasekera [24] presented a penalization method to simultaneously select significant variables and detect the true structure of partially linear varying coefficient models by local linear smoothing and the adaptive LASSO [34].

In this paper, we propose a new combined penalization technique to detect the true structure of the SIVCM, select significant variables and estimate the unknown index parameters and coefficient functions simultaneously. Specifically, we first use the B-spline basis to approximate the unknown coefficient functions of model (1.1). Then combining with the restraint $\|\beta\| = 1$, we adopt the "delete-one-component" method proposed by Yu and Ruppert [30] to construct the penalized least squares function. Note that, if the coefficient function $g_j(\cdot)$ has a zero derivative, then $g_j(\cdot)$ is a constant. Thus, based on the differentiability of B-spline basis functions, we penalize not only the coefficient functions but also their derivatives by using group SCAD penalty [31]. With proper choice of tuning parameters and knots, we show that the combined penalization model detection and variable selection procedure can consistently identify the true structure of the model, and the estimators of parametric and nonparametric components are consistent.

After the model detection step, the SIVCM (1.1) becomes a new semiparametric model—partially linear single-index varying coefficient model (PLSIVCM):

$$Y = \sum_{j=1}^{d} g_j(\tilde{\beta}^\tau \tilde{X})Z_j + \sum_{j=d+1}^{s} g_j Z_j + \varepsilon, \tag{1.2}$$

where $g_1(\cdot), \ldots, g_d(\cdot)$ are unknown functional coefficients, $g_{d+1}, \ldots, g_s$ $(s \le q)$ are unknown constant coefficients and $\tilde{\beta}$ is a ν-dimensional $(\nu \le p)$ unknown index parameter vector. Obviously, the classical partially linear single-index model (see [1,25]) is a special case of (1.2) with $d = 1$ and $Z_1 = 1$. We treat (1.2) as the final model. Compared with model (1.1), the appearance of linear part in (1.2) further reduces the complexity of the model and also aids to interpretation. Now we consider the estimation and inference for the parametric components and nonparametric functions of model (1.2). Note that the resulting penalized estimators are consistent, but their asymptotic distributions are not available. To overcome this, we update these estimators by the minimum average variance estimation method (MAVE, Xia et al. [27], Xia and Härdle [25]), and show that the resulting estimators are asymptotically normally distributed. The primary reason why we use MAVE is that, unlike most other methods, MAVE does not need to undersmooth the nonparametric function estimator to attain the $\sqrt{n}$-rate consistency for the parametric components.

The main innovations of the proposed method are as follows: Firstly, our method can select significant variables in the parametric and nonparametric components while identify the true model structure. We propose a new approach which directly shrinks the first derivative of the coefficient function to zero. This is very different from the method proposed by