Journal of Multivariate Analysis 125 (2014) 159–175

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Model structure selection in single-index-coefficient regression models

Zhensheng Huang [a,*], Zhen Pang [b], Bingqing Lin [b], Quanxi Shao [c]

[a] School of Science, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094, China
[b] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore
[c] CSIRO Computational Informatics, Private Bag 5, Wembley, WA 6913, Australia

ARTICLE INFO

Article history:
Received 1 April 2013
Available online 30 December 2013

AMS 2000 subject classifications:
primary 62G05
secondary 62G20

Keywords:
Delete-one-component method
Goodness-of-fit test
Penalized estimating equations (PEE)
Single-index models
Varying coefficient models
Variable selection

ABSTRACT

Single-index-coefficient regression models (SICRM) have been proposed and used in the literature for avoiding the "curse of dimensionality". However, there is no efficient model structure determination methodology for the SICRM. This may cause a tendency to use models that are much larger than required. In this paper, we propose a new procedure for model structure determination in the SICRM; that is, the penalized estimating equations (PEE) for variable selection that combines the "delete-one-component" method and the smoothly clipped absolute deviation penalty. The proposed PEE method can simultaneously identify significant variables of the index and estimate the nonzero coefficients of the index parameters. We also further study testing for nonparametric index-coefficient functions. Asymptotic properties for the proposed estimation procedure have been established. Under the appropriate conditions, we demonstrate that the proposed estimators have the oracle properties. Monte Carlo simulation studies are conducted to assess the finite sample performance of the proposed methods. A real example is analyzed as an illustration.

1. Introduction

Single-index models and varying coefficient models are powerful tools for dimension reduction and semiparametric modeling because they can effectively avoid "curse of dimensionality". Both models largely relax some restrictive assumptions on linear and nonparametric models. Specially, the single-index model is usually treated as the first step of the famous projection pursuit regression, which is proposed by [16,22,18,4] has also given a comprehensive discussion. The single-index model has also been extensively used in the projection pursuit regression, see [18,16]. In addition, its variety estimation methods have also been proposed in the literature. See [19,23,21] for more details and various applications. Varying coefficient models have wide applications in practice due to their flexibility since they were first proposed by Hastie and Tibshirani [20]. See [5,13,25,12,14,28,39,27] for various applications. But these two classes of models have their own limitations. For example, single-index models cannot reflect the additivity of covariates, while the performance of varying coefficient models can be poor if the varying coefficient contains moderate or high dimensional covariates.

* Corresponding author.
E-mail addresses: sta.zshuang@gmail.com, stahzs@126.com (Z. Huang).

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.12.006

As a compromise, we study a hybrid of single index and varying coefficient models, single-index-coefficient regression models (SICRM), which are written as

$$Y = g^T(\beta_0^T X)Z + \varepsilon, \tag{1}$$

where Y is the response variable, $X(\in R^d)$ and $Z(\in R^q)$ are the vectors of covariates, $g(\cdot) = (g_1(\cdot), \ldots, g_q(\cdot))^T$ is a q-dimensional unknown function vector. Generally, the first component of Z may be taken as 1. By properly choosing X and Z from all available predictors, particularly if we allow different choice of index in X for distinctive Z, the model can flexibly describe nonlinear mean structures of, and complex interactions for all predictors. To keep the structure simple, we focus on a single-index model as described hereafter. This model preserves sufficient flexibility yet would be quite practically feasible. Since each $g_j(\cdot)$ is unknown, the index $\beta_0^T X$ may be determined arbitrarily. For the sake of identifiability, we assume that $\|\beta_0\| = 1$ and the first component of β_0 is positive, where $\|\cdot\|$ denotes the Euclidean norm [3,49]. The model error ε has mean zero given X and Z.

The SICRM include many useful models such as: the single-index model with $Z = 1$ and $q = 1$ [46,40,26,44,19,23], the adaptive varying-coefficient linear model when $Z = X$ [11,31] and the varying-coefficient model if $d = 1$ and $\beta_0 = 1$ [37,43,47,2,20].

To the best of our knowledge, existing estimation procedures for model (1) were mainly built on the least-squared-based methods. Specifically, [42] was the first to study model (1) and used the kernel method to estimate the unknown function vector $g(\cdot)$ and obtain the estimators of β_0 by the least-squared method. Xue and Wang [45] considered the empirical likelihood confidence regions for the parametric component of this model. As a special example of model (1), [11] considered the adaptive varying-coefficient linear model and used the profile least-squares local approximation to estimate the parameter β_0. They also proposed to select locally significant variables by combining the T-statistic and the Akaike information criterion. The asymptotic theory of the procedure was later established by Lu et al. [31]. These methods may be adopted for estimation of models (1). However, a preliminary and natural question we first need to address is which covariates should be entered in the single index coefficient. It is well known that one may have a number of redundant covariates that increase the model complexity without significantly increasing the model accuracy.

This work was motivated by analyzing an environmental dataset consisting of four daily measurements of pollutants and two environmental factors, which was collected in Hong Kong between July 1, 1997 and December 31, 1997. The objective of the study is to examine the association between the levels of air pollutants and environmental factors and the sums of daily total hospital admissions for respiratory diseases. Xia [41] once used a multiple index model to examine a similar environmental dataset and advocated that the model is appropriate. However, to further improve the flexibility and interpretability of his models [41], we propose use of model (1) to explore the hidden structure in this type of dataset. Our numerical analysis presented in Section 6 shows promise of our idea.

In this context, we develop a methodology to identify and eliminate redundant single index covariates, and to further develop goodness-of-fit test and establish the corresponding asymptotic properties. The first aim is essentially a problem of variable selection, for which there is an extensive literature with several main approaches: the traditional methods such as the Bayesian information criterion, the Akaike information criterion and the bridge regression [15,34,1]; the least absolute shrinkage and selection operator (LASSO, [51,50,35]); the smoothly clipped absolute deviation (SCAD) penalty approach [9,32,30] and the minimum concave penalty [48]. In addition, there has been much work on variable selection of nonparametric and semiparametric models by using the regularization procedures including LASSO and SCAD. See, for example, [40,29,37,30,38]. See [10] for a fairly comprehensive survey.

Based on the SCAD penalty function, we may simultaneously select significant variables of the index and estimate the nonzero coefficient parameters in model (1). After simultaneously selecting the significant variables of $\beta_0^T X$ and estimating the nonzero coefficient parameters of β_0 in model (1), it is natural to consider the testing problem of the nonparametric parts. We mainly consider the goodness-of-fit test for the functional vector $g(\cdot)$ of model (1) here. Let $g(\beta_0^T X) = g_0(X; \beta_0)$ be a pre-specified working parameter model of which one would like to check its validity. We assume that $g_0(X; \beta_0)$ is a known finite-dimensional parametric functional vector, which includes a wide range of parametric types of regression models as special cases. This class of models includes the models with any shape constraints such as any combinations of the covariates X and the parameter β_0 of primary interests. This may improve the adaptability and interpretability of model (1). However, there has been no study focusing on this class of testing problems for the current model. In practice, establishing an efficient testing statistic is challenging for model (1) because of its complex structure. In this work, we develop an efficient testing procedure to deal with the proposed goodness-of-fit testing problem, which may also be used to select significant nonparametric components in model (1). The corresponding asymptotic properties are derived and the bootstrap testing procedure is also proposed.

The rest of this paper is organized as follows. In Section 2 we describe the regularized estimation procedure using the constraint of the parameters and the SCAD penalty and develop the corresponding consistency and oracle properties. We present the procedure of a goodness-of-fit test and establish its asymptotic properties in Section 3. The detailed computational algorithm is presented in Section 4. In Section 5, we illustrate the proposed methods using the simulated data examples. A real data example is given in Section 6. We then give a discussion in Section 7. The proofs of the main results are collected in the Appendix.