Computational Statistics and Data Analysis 65 (2013) 56–67

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Variable selection in high-dimensional partially linear additive models for composite quantile regression

Jie Guo[a], Manlai Tang[b,*], Maozai Tian[a], Kai Zhu[c]

[a] Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China
[b] Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
[c] National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 21 February 2012
Received in revised form 14 March 2013
Accepted 14 March 2013
Available online 2 April 2013

Keywords:
Adaptive Lasso
Composite quantile regression
High-dimension
Semiparametric additive partial linear model
Spline approximation
Variable selection

ABSTRACT

A new estimation procedure based on the composite quantile regression is proposed for the semiparametric additive partial linear models, of which the nonparametric components are approximated by polynomial splines. The proposed estimation method can simultaneously estimate both the parametric regression coefficients and nonparametric components without any specification of the error distributions. The proposed estimation method is empirically shown to be much more efficient than the popular least-squares-based estimation method for non-normal random errors, especially for Cauchy error, and almost as efficient for normal random errors. To achieve sparsity in high-dimensional and sparse additive partial linear models, of which the number of linear covariates is much larger than the sample size but that of significant covariates is small relative to the sample size, a variable selection procedure based on adaptive Lasso is proposed to conduct estimation and variable selection simultaneously. The procedure is shown to possess the oracle property, and is much superior to the adaptive Lasso penalized least-squares-based method regardless of the random error distributions. In particular, two kinds of weights in the penalty are considered, namely the composite quantile regression estimates and Lasso penalized composite quantile regression estimates. Both types of weights perform very well with the latter performing especially well in terms of precisely selecting significant variables. The simulation results are consistent with the theoretical properties. A real data example is used to illustrate the application of the proposed methods.

1. Introduction

Additive models have been proven to be very useful since they not only naturally generalize the linear regression models but also successfully circumvent the so-called 'curse of dimensionality' in nonparametric regression (see, Breiman and Friedman (1985) and Hastie and Tibshirani (1986, 1990)). Additive partial linear models (APLMs) can be considered as a modification of additive models with a parametric (linear) part or as a nontrivial generalization of the linear model with additive components. In other words, APLMs allow the response variable to depend linearly on some covariates and nonlinearly on the remaining variables. Suppose that $\{(\mathbf{X}_i, \mathbf{Z}_i, Y_i) : i = 1, \ldots, n\}$ is an independent and identically distributed (i.i.d.) sample from $(\mathbf{X}, \mathbf{Z}, Y)$, where Y is a response variable, $\mathbf{X} = (X_1, \ldots, X_d)^T$ is a d-dimensional covariate

[*] Corresponding author. Tel.: +852 3411 7018; fax: +852 3411 5811.
E-mail addresses: guojieruc2009@ruc.edu.cn (J. Guo), mltang@math.hkbu.edu.hk (M. Tang), mztian@ruc.edu.cn (M. Tian), zhukai@bao.ac.cn (K. Zhu).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.03.017

vector and $\mathbf{Z} = (Z_1, \dots, Z_p)^T$ is another p-dimensional covariate vector. The semiparametric APLM is defined as follows:

$$Y_i = \mathbf{X}_i^T \alpha + \sum_{j=1}^{p} g_j(Z_{ij}) + \varepsilon_i, \tag{1.1}$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{id})^T \in \mathbb{R}^d$, $\alpha = (\alpha_1, \dots, \alpha_d)^T$ is a vector of corresponding regression coefficients, $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{ip})^T \in \mathbb{R}^p$ with Z_{ij} being the jth component of $\mathbf{Z}_i$, g_j's are one-dimensional unknown smooth functions of Z_{ij}, and $Y_i \in \mathbb{R}$. We assume that ε_i's are i.i.d. random errors which are independent of $(\mathbf{X}_i, \mathbf{Z}_i)$'s. To ensure identifiability of the nonparametric functions, we assume that $Eg_j(Z_j) = 0$ for $j = 1, \dots, p$.

For APLM estimation, one can consult Stone (1985) and Opsomer and Ruppert (1997) for backfitting algorithm, Fan et al. (1998), Opsomer and Ruppert (1999) and Liang et al. (2008) for methods based on kernel regression, and Koenker (2011) for quantile regression based method. However, all these methods are not able to estimate the parameters and nonparametric components simultaneously, and the existing estimation methods are all based on least squares approach and the assumptions that the error is normally distributed or has finite variance, which are not always true in most applications.

In this paper, we first propose to use polynomial splines to approximate the nonparametric components in model (1.1). With this approximation, each nonparametric component is represented by a linear combination of spline basis functions. Consequently, the estimation problem of model (1.1) becomes a problem of estimating the coefficients in the linear combinations and the parameters and nonparametric components can thus be estimated simultaneously. Next, we propose to apply the composite quantile regression (CQR) approach to estimate the coefficients due to its appealing properties and performance (see, Zou and Yuan (2008)). It has been shown that the relative efficiency of the CQR estimator to the least squares estimator is greater than 70% regardless of the error distribution. Moreover, the CQR estimator could be much more efficient and sometimes arbitrarily more efficient than the least squares estimator for non-normal errors, especially for Cauchy errors, and almost as efficient for normal errors (see, Guo et al. (2012)).

In practice, a large number of variables may be collected and some of them are insignificant and should be excluded from the final model. There has been active methodological research in penalized methods for significant variable selection in linear parametric models. Examples include the bridge regression (Frank and Friedman, 1993), the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the elastic net penalty (Zou and Hastie, 2006) and the adaptive Lasso (Zou, 2006). However, variable selection in the linear part of APLMs is much more difficult and challenging and scattered studies have been found in statistics literature. Liu et al. (2011) proposed a SCAD penalized method based on spline approximation in APLMs. Wei (2012) proposed a two-step adaptive group Lasso penalized method for group variable selection based on spline approximation. Marra and Wood (2011) introduced two very simple but effective shrinkage methods and an extension of the nonnegative garrote estimators for the problem of variable selection within the class of generalized additive models. Du et al. (2012) introduced a penalty that combines the adaptive empirical L_2-norms of the nonparametric component functions and the SCAD penalty on the coefficients in the parametric part to achieve simultaneous model selection for both nonparametric and parametric parts in semiparametric regression models with additive nonparametric components and high dimensional parametric components under sparsity assumptions.

To select significant variables in the linear part of high-dimensional and sparse APLMs, we use the adaptive penalized CQR approach by considering two kinds of weights in the penalty. The Type I weight is constructed using the corresponding CQR estimates (which is similar to the idea proposed in Zou and Yuan (2008)) while the Type II weight is constructed using the Lasso penalized CQR estimates, which has not been considered in literatures yet. We find that the adaptive penalized CQR approach incorporated with the Type II weight performs extraordinarily well, especially in terms of precisely selecting the significant variables. The proposed variable selection procedure enjoys the oracle properties and works very well regardless of the error distribution and choice of weight.

The rest of the paper is organized as follows. Section 2 studies the proposed estimation and variable selection procedures, respectively, based on CQR of high-dimensional and sparse APLMs, and the theoretical properties of the estimators and the oracle properties of the variable selection procedure are also presented. Numerical comparisons and simulation studies are described in Section 3. Section 4 illustrates the application of the proposed methods with a real dataset. Proofs of the technical results are presented in the Appendix.

2. Estimation and variable selection

2.1. Spline approximation and estimation

Suppose that each Z_j takes values in a compact support $[a, b]$, where $a < b$ are finite numbers. Polynomial splines are piecewise polynomials connected smoothly over a set of interior points or knots and we allow the number of knots to increase with the sample size. For each $1 \le j \le p$, let $\xi_j = \{a = \xi_{j0} < \xi_{j1} < \cdots < \xi_{jA_j} < \xi_{A_j+1} = b\}$ be knot sequences with A_j interior knots on $[a, b]$. For simplicity, we use the same knot sequence for all $j = 1, \dots, p$; that is, $\xi = \{a = \xi_0 < \xi_1 < \cdots < \xi_A < \xi_{A+1} = b\}$. Here, $[a, b]$ is divided into $A + 1$ subintervals $I_{A_t} = [\xi_t, \xi_{t+1})$, $t = 0, \dots, A - 1$ and $I_{AA} = [\xi_A, \xi_{A+1}]$, where $A \equiv A_n = n^{\upsilon}$ with $0 < \upsilon < 0.5$ is a positive integer such that $\max_{1 \le k \le A+1} |\xi_k - \xi_{k-1}| = O(n^{-\upsilon})$ (see, Huang et al. (2010)).