Journal of Multivariate Analysis 141 (2015) 81–103

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Spline estimator for simultaneous variable selection and constant coefficient identification in high-dimensional generalized varying-coefficient models

Heng Lian [a], Jie Meng [b], Kaifeng Zhao [c,*]

[a] School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, 2052, Australia
[b] Department of Marketing and Management, Macquarie University, North Ryde, NSW, 2109, Australia
[c] Department of Mathematics, School of Science, Harbin Institute of Technology, Harbin, 150001, PR China

ARTICLE INFO

Article history:
Received 2 October 2014
Available online 24 June 2015

AMS subject classification:
62G20

Keywords:
B-spline basis
Diverging parameters
Group lasso
Quasi-likelihood

ABSTRACT

In this paper, we are concerned with two common and related problems for generalized varying-coefficient models, variable selection and constant coefficient identification. Starting with a specification of generalized varying-coefficient models assuming possible nonlinear interactions between the index variable and all other predictors, we propose a polynomial-spline based procedure that simultaneously eliminates irrelevant predictors and identifies predictors that do not interact with the index variable. Our approach is based on a double-penalization strategy where two penalty functions are used for these two related purposes respectively, in a single functional. In a "large p, small n" setting, we demonstrate the convergence rates of the estimator under suitable regularity assumptions. Based on its previous success on parametric models, we use the extended Bayesian information criterion (eBIC) to automatically choose the regularization parameters. Finally, post-penalization estimator is proposed to further reduce the bias of the resulting estimator. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed procedures and an application to a leukemia dataset is presented.

1. Introduction

Generalized linear models (GLM) provide an extension of linear models in dealing with different types of responses, including for example binary data and count data [24]. However, such parametric models are not flexible enough to capture the true underlying relationships between covariates and responses. Of particular interests to us in this paper is the generalized varying-coefficient models (GVCM) [13,3]. Let Y be a response variable and (X, T) is the associated covariates where T is one dimensional and $X = (X_1, \ldots, X_p)^T$. The (conditional) mean of the response, $\mu = E[Y|X, T]$, takes the form

$$g(\mu) = X^T\alpha(T), \tag{1}$$

where $\alpha(T) = (\alpha_1(T), \ldots, \alpha_p(T))^T$. The index variable T is usually some variable related to time or age in many applications whose interactions with other predictors are believed to be of importance. Meanwhile, we assume the conditional variance $\mathrm{Var}(Y|X, T) = V(\mu)$ only depends on the conditional mean.

* Corresponding author.
E-mail address: kaifengzhao66@hotmail.com (K. Zhao).

http://dx.doi.org/10.1016/j.jmva.2015.06.011
0047-259X/

Recently, semiparametric regression models receive increasing attention as they represent an appropriate trade-off of complexity with simplicity between parametric and nonparametric modeling. Here the predictors are partitioned into two subsets, $X^{(1)} = (X_1, \ldots, X_{p_1})$ and $X^{(2)} = (X_{p_1+1}, \ldots, X_p)$, and (1) is replaced by

$$g(E[Y|X, T]) = g(\mu) = X^{(1)T}\alpha(T) + X^{(2)T}\beta, \tag{2}$$

where now only some of the predictors exhibits interactions with T [10,19]. If the predictors associated with varying and non-varying coefficients could be correctly specified, for example, based on some expert knowledge, then the semiparametric version is the clear choice. However, such prior information is difficult to obtain in applications. Thus one of our goals is to automatically produce the semiparametric model even when no prior information is available for correct specification of (2).

Identifying constant coefficients in GVCMs has indeed received attention previously and either cross-validation [33] or generalized likelihood ratio test [3,17,10] can be used. We develop in this paper a penalization based approach that automatically separates constant coefficient from varying ones. When the number of predictors is large, penalization based approach is more computationally efficient [20].

High-dimensionality is an important characteristic of many modern datasets. Our investigation is motivated by a ALL study [6] investigating genetic mechanisms based on microarray assays from leukemia patients. Of particular interest is the classification of disease/normal samples. Taking the probesets expression levels as covariates X and the age of the patient as T, a logistic varying coefficient models with $g(\mu) = \log \mu/(1-\mu)$ and $V(\mu) = \mu(1-\mu)$ is a suitable choice. We are interested in identifying genes related to disease status and in the mean time aim to identify a more parsimonious partially linear structure that allows more efficient estimation.

Traditional variable selection methods such as stepwise regression and best subset selection is computationally infeasible when the number of predictors is large, as argued in [20], this is part of the reason why penalization based method [30,11,37,16,23] has gained popularity in recent years. This motivates us to develop a penalization based approach for both variable selection and constant coefficient identification in a consistent framework. Starting with a nonparametric varying-coefficient model (1) which includes a large number of predictors, our approach automatically produces as a final output a semiparametric model (2) with a small number of predictors. Such approach for GVCPLMs is not available in the current literature. In addition, high-dimensionality ($p \gg n$) poses serious theoretical challenges.

This paper is organized as follows. In Section 2, we propose a penalization procedure that achieves both variable selection and constant coefficient identification simultaneously. Unlike Lam and Fan [19] which is based on local polynomial regression, we use polynomial splines to approximate the nonparametric coefficients, which is computationally easier since it directly reduces the nonparametric model to parametric GLM as far as computations are concerned. Two regularization parameters are automatically chosen using extended Bayesian information criterion (eBIC) [4,5]. To reduce bias resulting from penalization, we also use an unpenalized version of the functional to obtain our final estimator. In Section 3, Monte Carlo simulation studies are carried out for both Poisson regression and logistic regression models to demonstrate the performance of the proposed method. In addition, a real dataset is used as an illustration. The paper concludes in Section 4 and the Appendix A contains all technical proofs, while Appendix B contains discussions on some complicated eigenvalue assumptions.

Zhang et al. [35]; Hu and Xia [15]; Tang et al. [29]; Noh and Van Keilegom [25] have considered partially linear structure identification in both additive and varying-coefficient models. However, these works focus only on the fixed p case and it remains very challenging in high-dimensional situations. Our main contribution is to develop the methodology and theory in high-dimensional settings, as motivated by the genetic data. The partially linear structure makes the model flexible and parsimonious at the same time. The high-dimensional additive model is considered in [22] based on adaptive lasso penalties. Besides that we consider the varying coefficient model instead of the additive model here, the generalized version can deal with both count responses and binary responses and thus the extension is significant.

2. Spline estimator and sampling properties

The data we observe for the ith subject or unit are (X_i, T_i, Y_i), $i = 1, \ldots, n$, where $X_i = (X_{i1}, \ldots, X_{ip})^T$ are the predictors and T_i is the index variable. The true model is assumed to be that of (2) where $X^{(1)} = (X_1, \ldots, X_{p_1})$ are the predictors associated with truly varying coefficient, $X^{(2)} = (X_{p_1+1}, \ldots, X_{p_1+p_2})$ are the predictors associated constant coefficients, thus the total number of nonzero coefficients is $s = p_1 + p_2$ and the rest $p - s$ predictors do not appear in the true model. We remind the readers that the information regarding which coefficients are constant, varying or zeros is not made known to the statistician working with the model. However, without loss of generality, we assume that in the true model the first p_1 predictors are associated with varying coefficients, and the next p_2 associated with nonzero constant coefficients. We denote the true coefficient vectors by $\alpha_0 = (\alpha_{01}, \ldots, \alpha_{0p_1})^T$ and $\beta_0 = (\beta_{0,p_1+1}, \ldots, \beta_{0s})^T$. For simplicity and as usually assumed in the literature, the index variable T is univariate with a distribution supported on the interval [0, 1]. The extension to multi-dimensional T is possible but rarely used in practice due to "curse of dimensionality".

2.1. Estimation procedure based on adaptive group lasso

The (negative) quasi-likelihood function is defined by

$$Q(\mu, y) = \int_{\mu}^{y} \frac{y - s}{V(s)}\, ds,$$