Journal of Multivariate Analysis 154 (2017) 40–58

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Quantile index coefficient model with variable selection

Weihua Zhao [a], Heng Lian [b,*]

[a] School of Science, Nantong University, Nantong, 226019, PR China

[b] Department of Mathematics, City University of Hong Kong, Kowloon Tong, HK, 999077, Hong Kong

ARTICLE INFO

Article history:
Received 11 April 2016
Available online 31 October 2016

AMS subject classifications:
62G08
G2G20

Keywords:
Asymptotic normality
B-splines
Check loss minimization
Mixing condition
Variable selection

ABSTRACT

We consider conditional quantile estimation in functional index coefficient models for time series data, using regression splines, which gives more complete information on the conditional distribution than the conditional mean model. An important technical aim is to demonstrate the faster rate and asymptotic normality of the parametric part, which is achieved through an orthogonalization approach. For this class of very flexible models, variable selection is an important problem. We use smoothly clipped absolute deviation (SCAD) penalty to select either the covariates with functional coefficients, or covariates that enter the index, or both. We establish the oracle property of the penalization method under strongly mixing (α-mixing) conditions. Simulations are carried out to investigate the finite-sample performance of estimation and variable selection. A real data analysis is reported to demonstrate the application of the proposed methods.

1. Introduction

Linear time series models are well developed in the statistics and econometrics literature, including the well-known ARMA models. However, linear models cannot capture some nonlinear effects often seen in real data, as has been highlighted in [8,13,19]. Even with parametric nonlinear models, the pre-specified form of nonlinearity is too stringent in many applications. Therefore, considering the curse of dimensionality in full-blown nonparametric models, semiparametric modeling has become popular in recent years [2,29,40].

One popular class of models for semiparametric modeling is the functional coefficient model (FCM) introduced in [7,21], for cross-sectional data and time series data, respectively. The model is given by

$$Y_i = \sum_{j=1}^{p} g_j(X_i) Z_{ij} + e_i,$$

given observations $Y_i, X_i, \mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top, i = 1, \ldots, n$, where X is usually called an index variable in this context. The FCM generalizes the linear models by allowing the regression coefficient to depend smoothly on the index variable, e.g., time. FCM has been widely studied in the literature; see [2,18,19,24,28,36].

If the index X_i is a multidimensional vector, it is hard to fit the FCM directly since accurate estimation of multivariate smooth functions typically requires an exorbitantly large sample size. Xia and Li [49] proposed an elegant solution for multivariate $\mathbf{X}$ by introducing an index structure to effectively transform the index vector into a one-dimensional index

* Corresponding author.
E-mail address: heng.lian@unsw.edu.au (H. Lian).

http://dx.doi.org/10.1016/j.jmva.2016.10.009
0047-259X/

variable as in

$$Y_i = \sum_{j=1}^{p} g_j(\mathbf{X}_i^\top \boldsymbol{\beta}) Z_{ij} + e_i. \tag{1}$$

Compared to FCM, the extra complication is to estimate simultaneously the index coefficient $\boldsymbol{\beta}$ together with smooth functions $\mathbf{g} = (g_1, \ldots, g_p)^\top$. Fan et al. [17] proposed efficient estimation methods and used a stepwise method combined with Akaike's information criterion (AIC, [1]) for selecting significant covariates $\mathbf{X}$ in the index. Lu et al. [38] provided some asymptotic theory for estimation. Cai et al. [3] used a smoothly clipped absolute deviation (SCAD) penalty to select the significant covariates in both $\mathbf{X}$ and $\mathbf{Z}$.

In a seminal paper, Koenker and Bassett [32] proposed linear quantile regression to examine the effects of an observable covariate on the distribution of a dependent variable, with special interest in the tail of the distribution. Since then, quantile regression has been widely used in various disciplines, including finance, economics, medicine, and biology; see for example the popular monograph [31]. Parallel to linear mean regression based on least squares, linearity in quantile regression has been relaxed to accommodate possibly nonlinear effects in nonparametric and semiparametric quantile regression models. This large literature includes [22] for spline estimation of nonparametric quantile regression, [6,20,52] for local polynomial estimation, [34] for partial linear models, [11,25,37] for additive models, [4,5,46] for functional coefficient models, and [30,33,48] for single-index models.

In this paper, we are interested in *quantile* regression for (1), which we call the functional index model (FIM), and we are in particular interested in the associated variable selection problem, which was not considered before. Recent challenging topics in statistics include the development of automatic variable selection procedures intended to find the relevant parameters among all candidate parameters and simultaneously estimate them. As argued in [36], traditional variable selection methods, such as stepwise regression and best subset selection, are computationally infeasible when the number of predictors is large, and this is part of the reason why the penalization based method has gained popularity in recent years.

Substantial progress has been made on the problem of variable selection for linear models and generalized linear models [9,14,15,26,41,53,55,56]. More recently, variable selection methods using penalty functions in nonparametric or semiparametric settings have been developed. For example, Xie and Huang [50] developed variable selection based on penalization for partially linear models, and additive models were investigated in [27,51]. For semiparametric functional coefficient models, Li and Liang [36] used penalization to select the significant predictors in the parametric components, while a group penalization method for selecting nonparametric functions was proposed in [45,42]. These previous works motivated us to develop a penalization based approach for variable selection in FIM for both the index parameters and the smooth functions.

The papers mentioned above on variable selection are based on the assumption that the observed data are independent and identically distributed (i.i.d.). Under a non i.i.d. setting, the studies on penalized variable selection are scarce; see, e.g., [3,43]. In this paper, we will develop theory and methodology for the quantile FIM using polynomial spline estimation, under strongly mixing assumptions which are more appropriate for financial or other data with observations that are time dependent. Polynomial spline estimation provides an alternative to local polynomial estimation method. The comparative advantages of spline methods were carefully documented in [35], among which the most notable is the computational convenience, as argued in [35,44].

The rest of the paper is organized as follows. In Section 2, we present the estimation method using polynomial splines, and asymptotic properties of the estimators are considered. Then we further consider variable selection for both the index parameter and the smooth functions. Section 3 presents some Monte Carlo studies of the finite-sample performance of the estimators, as well as an empirical application of the method. We end the paper with a short discussion in Section 4. The technical proofs are relegated to the Appendix.

2. Quantile index coefficient models

2.1. Estimation methods without variable selection

We consider the index coefficient model

$$Y_i = \mathbf{g}^\top(\mathbf{X}_i^\top \boldsymbol{\beta})\mathbf{Z} + e_i,$$

where $(\mathbf{X}_i, \mathbf{Z}_i, Y_i, e_i)$ is strictly stationary, $\mathbf{g}(\cdot) = (g_1(\cdot), \ldots, g_p(\cdot))^\top$ are the p coefficient functions whose argument is the index $\mathbf{X}_i^\top \boldsymbol{\beta}$, $\Pr(e_i \le 0 | \mathbf{X}_i, \mathbf{Z}_i) = \tau$, and $\mathbf{X}_i$ and $\mathbf{Z}_i$ are q-dimensional and p-dimensional covariates, respectively. We assume the first component of $\mathbf{Z}$ is 1 and thus no separate intercept is explicitly written. We also allow common variables in $\mathbf{Z}$ and $\mathbf{X}$; in particular, it is permissible that $\mathbf{Z} = (1, \mathbf{X}^\top)^\top$. In that case, we assume that the identifiability conditions stated in Theorem 1 of [17] are satisfied. Besides, we always assume $\|\boldsymbol{\beta}\| = 1$ and $\beta_1 > 0$.

We use polynomial splines to approximate the components. Let $\tau_0 = a < \tau_1 < \cdots < \tau_{K'} < b = \tau_{K'+1}$ be a partition of $[a, b]$ into subintervals $[\tau_k, \tau_{k+1})$, $k = 0, \ldots, K'$ with K' internal knots. We only restrict our attention to equally spaced knots although data-driven choices can be considered such as putting knots at certain sample quantiles of the observed covariate