Journal of Multivariate Analysis 141 (2015) 49–66

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Quantile regression for dynamic partially linear varying coefficient time series models

Heng Lian

School of Mathematics and Statistics, University of New South Wales, Sydney, 2052, Australia

ARTICLE INFO

Article history:
Received 20 March 2014
Available online 26 June 2015

AMS subject classification:
62G20

Keywords:
Autoregressive models
Model structure recovery
SCAD penalty
Schwarz information criterion (SIC)
Splines

ABSTRACT

In this article, we consider quantile regression method for partially linear varying coefficient models for semiparametric time series modeling. We propose estimation methods based on general series estimation. We establish convergence rates of the estimator and the root-n asymptotic normality of the finite-dimensional parameter in the linear part. We further propose penalization-based method for automatically specifying the linear part of the model as well as performing variable selection, and show the model selection consistency of this approach. We illustrate the performance of estimates using a simulation study.

1. Introduction

Various nonparametric and semiparametric models have been developed for dynamic time series data analysis. One popular semiparametric model that represents a good balance between flexibility and parsimony is the varying coefficients model, which has found many applications in finance, economics, medicine and biology [16,1,34,12,13,29]. These models extend the classical linear models by allowing the coefficients to depend smoothly on an index variable. Many estimation procedures have been proposed previously for independent data [11,12,6,18,19,7]. Cai et al. [2] considered estimation based on kernel methods for time series data, and Cai and Li [3] further extended this to panel data model.

The parametric quantile regression introduced by Koenker and Bassett [22] has been well developed in the econometrics and statistics literature. When the distribution of the errors in the model is heavy tailed or the data contain some outliers, it is well known that median regression, a special case of quantile regression, is more robust than mean regression. More importantly, it can be used to obtain a large collection of conditional quantiles to characterize the entire conditional distribution. To construct a richer class of regression models capturing flexibly the relationships between the covariates and the response distribution, nonparametric quantile estimation has been studied in [15,35]. For varying coefficient models, Kim [21] studied quantile regression for independent data using splines, and Cai and Xu [5] used local polynomial estimation method for time series data.

The varying coefficient models, although more parsimonious than the fully nonparametric models, can still overfit the data when some covariate effects are actually linear, which motivated the partially linear varying coefficient model (PLVC) studied in [32] for independent data using splines and Cai and Xiao [4] for dynamic time series data using local polynomials.

The present article will develop theory and methodology for analyzing stationary time series data in the quantile PLVC model using general series estimation methods. Series estimation methods provide an alternative to local polynomial

E-mail address: heng.lian@unsw.edu.au.

http://dx.doi.org/10.1016/j.jmva.2015.06.013
0047-259X/

estimation method. The comparative advantages of series estimation methods were carefully documented in [23], among which the most notable is the computational convenience, although it is not our main intention here to promote series estimation methods. The disadvantage is that the exact bias term for the nonparametric part is not known in the literature making demonstration of asymptotic normality for the nonparametric part difficult. We here use series estimation method mainly due to personal taste, and also due to that asymptotic properties of semiparametric PLVC quantile regression on time series data using this method has not been treated, which is much more complicated than the independent case (asymptotic properties using the local polynomial estimation methods have been established in [4]).

Moving one step further, we also consider penalization-based variable selection for PLVC quantile regression, which has not been considered in the context of time series data. Variable selection for some common time series models, for example for autoregressive models, has the favorable effect of determining the order of the model automatically, and simultaneously with estimation. Since there is an enormous literature on penalization-based variable selection, we just list a small number of them here including [26,8,9,27,30,31,24,17], which considered both parametric and semiparametric models, with apologies to those whose works are missed.

For the PLVC model, a plaguing problem is to determine which covariates have linear effect for correct model specification. Motivated by Lian [25] which studied partially linear additive quantile regression for independent data using splines, we consider the problem of automatic partially linear structure discovery for dynamic varying coefficient time series quantile regression. We demonstrate that we can still harvest the advantages of the more efficient parametric estimation for some covariates (if their effects are indeed linear) without having to specify the linear part before estimation. Instead, the linear part will emerge automatically as a by-product of statistical estimation. In the framework of quantile regression, the type of effect for a specific covariate can vary at different quantiles. That is, some covariate might have nonlinear effect at one quantile level but linear effect at another level.

The rest of the article is organized as follows. In Section 2, we formally present the partially linear functional coefficient model, the estimation procedure, and statistical properties. In Section 3, we consider the case where the linear part is not specified a priori and use penalized estimation to identify the model structure. We show the oracle property of the penalized estimator. In other words, this method estimates the irrelevant coefficients as zero, and the non-varying coefficients as nonzero constant, with probability approaching one. Convergence rates of the nonparametric part is also established. In Section 4, we briefly discuss computational aspects and present some numerical examples for finite sample performance. Section 5 presents some concluding remarks. All technical proofs are relegated to the Appendix.

2. Spline estimation of partially linear functional coefficient models

Let $(\mathbf{X}_i, \mathbf{Z}_i, U_i, Y_i), i = 1, \ldots, n$ be jointly stationary processes. At a given quantile $\tau \in (0, 1)$, we assume the PLVC quantile regression model

$$Y_i = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_\tau(U_i) + \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\alpha}_\tau + \epsilon_{\tau i},$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip_1})^{\mathrm{T}}$ is p_1-dimensional, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip_2})^{\mathrm{T}}$ is p_2-dimensional, $P(\epsilon_{\tau i} \le 0 | \mathbf{Z}_i, \mathbf{X}_i, U_i) = \tau$, and U_i is called the smoothing variable or the index variable, which might or might not be one component of $(\mathbf{X}_i, \mathbf{Z}_i)$. We only consider one-dimensional smoothing variable $U_i \in R$. Although multi-dimensional U_i can possibly be accommodated, in practice this is rare due to the worry of curse of dimensionality in high dimensional nonparametric regression. We assume the smoothing variable U_i takes values in a bounded interval $[-T, T]$, which is typical in series estimation methods. Since we consider a fixed quantile level, in the following the subscript τ will be omitted from the notations.

To compute the quantile regression estimator, we use a linear combination of known basis functions (such as power series, splines, Fourier series, wavelets, etc.), with the property that the linear combination can approximate the coefficients $\beta_j(u), j = 1 \ldots, p_1$ well (with more rigorous assumptions presented below). Without loss of generality, we assume that the constant function is in the space spanned by the basis functions. For simplicity of notation, we assume that all p_1 coefficients are approximated by the same set of basis functions. Theoretically it is straightforward to extend to the more general case where a set of p_1 different bases are used. However, in practice this is rarely done due to the difficulty of choosing multiple bases in a principled way. Let $p^K(u) = (p_1^K(u), \ldots, p_K^K(u))^{\mathrm{T}}$ be the sequence of basis functions and the coefficients are approximated by $\beta_j \approx \boldsymbol{\gamma}_j^{\mathrm{T}} p^K$ with $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jK})^{\mathrm{T}}$.

To estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_{p_1}^{\mathrm{T}})^{\mathrm{T}}$, we minimize

$$Q(\boldsymbol{\gamma}, \boldsymbol{\alpha}) := \sum_{i=1}^{n} \rho(Y_i - p^K(\mathbf{X}_i, U_i)^{\mathrm{T}} \boldsymbol{\gamma} - \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\alpha}), \tag{1}$$

where $\rho(x) = x(\tau - I\{x < 0\})$ is the check loss and $p^K(\mathbf{X}_i, U_i) = (X_{i1} p^K(U_i)^{\mathrm{T}}, \ldots, X_{ip_1} p^K(U_i)^{\mathrm{T}})^{\mathrm{T}}$. Note this is a one-step estimation procedure, as opposed to the two-step procedure used in [4] in order to achieve root-n consistency of the linear part.

Denoting the minimizer by $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\alpha}})$, we estimate $\beta_j(u)$ by

$$\hat{\beta}_j(u) = \sum_{k=1}^{K} \hat{\gamma}_{jk} p_k^K(u).$$