Computational Statistics and Data Analysis 118 (2018) 66–83

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimation and testing for time-varying quantile single-index models with longitudinal data

Jianbo Li [a], Heng Lian [b,*], Xuejun Jiang [c], Xinyuan Song [d]

[a] School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, China
[b] Department of Mathematics, City University of Hong Kong, Hong Kong
[c] South University of Science and Technology, Shenzhen, China
[d] Department of Statistics, Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 18 February 2017
Received in revised form 8 August 2017
Accepted 14 August 2017
Available online 18 September 2017

Keywords:
Asymptotic normality
B-splines
Check loss minimization
Single-index models
Quantile regression

ABSTRACT

Regarding semiparametric quantile regression, the existing literature is largely focused on independent observations. A time-varying quantile single-index model suitable for complex data is proposed, in which the responses and covariates are longitudinal/functional, with measurements taken at discrete time points. A statistic for testing whether the time effect is significant is developed. The proposed methodology is illustrated using Monte Carlo simulation and empirical data analysis.

1. Introduction

Single-index models (SIMs) provide an efficient means of coping with high-dimensional nonparametric estimation problems (Haerdle et al., 1993; Yu and Ruppert, 2002) and avoiding the "curse of dimensionality" in nonparametric problems by assuming that the response is only related to a single linear combination of the covariates. The fitting of SIMs, commonly based on splines or kernel methods, is widely applied in the literature. Nevertheless, efficient and stable estimation of SIMs remains a challenging problem that has inspired many recent works in this area (Wang and Yang, 2009; Bai et al., 2009; Wang et al., 2010; Li et al., 2010; Liang et al., 2010; Choi et al., 2011; Gramacy and Lian, 2012).

However, the restriction of these works on mean regression, that is, on estimating the conditional mean regression function, may be a limitation. As a useful supplement to mean regression, quantile regression (Koenker and Bassett Jr, 1978; Koenker, 2005) produces a more complete description of the conditional response distribution and is more robust to heavy-tailed random errors. In particular, quantile regression can uncover different structural relationships between covariates and responses at the upper or lower tails, and these relationships are occasionally of significant interest in econometrics and biomedical applications. This interest inspired a series of works on quantile SIMs recently (Wu et al., 2010; Kong and Xia, 2012). Another importance piece of work is Ma and He (2016) which considered quantile SIM using splines and demonstrated the effectiveness of the profiled approach.

Quantile SIMs assume that $Q_{y|\mathbf{x}}(\tau) = g(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}(\tau))$, where $Q_{y|\mathbf{x}}$ denotes the conditional τ-th quantile of a response y, g is an unknown link function, $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ is the covariate vector, and $\boldsymbol{\beta}(\tau)$ is the index parameter that depends on the quantile level $\tau \in (0, 1)$, but we omit this dependence in our notations in the following. The aforementioned studies are appropriate

* Corresponding author.
E-mail address: henglian@cityu.edu.hk (H. Lian).

http://dx.doi.org/10.1016/j.csda.2017.08.011
0167-9473/

for cross-sectional data in which the observations are assumed independent. To deal with data involving subjects that are repeatedly measured at different time points, a straightforward extension without additional efforts is considering models of the form $Q_{y(t)|\mathbf{x}(t)}(\tau) = g(\mathbf{x}^{\mathrm{T}}(t)\boldsymbol{\beta})$. However, in this form, the effect of time on the response is only through the values of the predictors at that time because the link function is time independent. This can be too restrictive in certain applications. Our objective is to extend the quantile SIMs to longitudinal/functional data using the following model structure:

$$Q_{y(t)|\mathbf{x}(t)}(\tau) = g(\mathbf{x}^{\mathrm{T}}(t)\boldsymbol{\beta}, t).$$

This specification allows for longitudinal data with both responses and covariates depending on the time variable. More importantly, the effect of the index $\mathbf{x}(t)^{\mathrm{T}}\boldsymbol{\beta}$ can be time varying through the bivariate link function g, which also depends on time. Such structure was previously attempted for mean regression in Jiang and Wang (2011).

The price to pay for a flexible time-varying model structure is that it loses efficiency when the effect of the index is actually not time varying. Thus, for data analysis, a hypothesis test should be conducted to determine whether the time effect is significant. If the test do not reject the null hypothesis that the link function is time independent, then a non-time-varying model can be refitted. We propose a rank score test for this purpose.

The rest of the article is organized as follows. Section 2.1 considers the model and estimation for the time-varying single-index quantile regression, and Section 2.2 establishes the asymptotic properties. Section 3 tests whether the link function is time varying. Section 4 presents specific Monte Carlo studies and Section 5 contains an empirical analysis of the AIDS data. Section 6 concludes with a discussion. Appendix contains all of the technical proof details.

2. Time-varying quantile single-index models

2.1. Model and estimation

In longitudinal data setting, the responses are observed at discrete time points. Assuming for the ith subject, $1 \le i \le n$, we make observations at time points $t_{ij}, j = 1, \ldots, m_i$, the dynamic semiparametric quantile regression model is written as

$$y_{ij} = g(\mathbf{x}_{ij}{}^{\mathrm{T}}\boldsymbol{\beta}, t_{ij}) + e_{ij},$$

where $y_{ij} = y(t_{ij})$ is the response, $\mathbf{x}_{ij} = \mathbf{x}(t_{ij})$ is the p-dimensional covariates, g is an unknown bivariate link function, and e_{ij} satisfies $P(e_{ij} \le 0|\mathbf{x}_{ij}) = \tau$. For identifiability, we assume $\|\boldsymbol{\beta}\| = 1$ and its first component is positive. We use $\boldsymbol{\beta}_0$ to denote the true value of $\boldsymbol{\beta}$ and use simply g (without subscript 0) to denote the true link function.

To take into account the unit norm constraint, we use the popular "delete-one-component" method (Yu and Ruppert, 2002; Cui et al., 2011). We can write $\boldsymbol{\beta} = ((1 - \|\boldsymbol{\beta}^{(-1)}\|^2)^{1/2}, \beta_2, \ldots, \beta_p)^T$, where $\boldsymbol{\beta}^{(-1)} = (\beta_2, \ldots, \beta_p)^T$ is $\boldsymbol{\beta}$ without the first component. Thus, $\boldsymbol{\beta}$ is a function of $\boldsymbol{\beta}^{(-1)}$. The $p \times (p-1)$ Jacobian matrix is

$$\mathbf{J} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\beta}^{(-1)}} = \begin{pmatrix} -\dfrac{\boldsymbol{\beta}^{(-1)}}{(1 - \|\boldsymbol{\beta}^{(-1)}\|^2)^{1/2}} \\ \mathbf{I}_{p-1} \end{pmatrix},$$

where $\mathbf{I}_{p-1}$ is the $(p-1) \times (p-1)$ identity matrix. In the rest of the paper, $\mathbf{J}$ is always evaluated at the true $\boldsymbol{\beta}_0$.

Assuming $\mathbf{x}_{ij}{}^{\mathrm{T}}\boldsymbol{\beta}$ is supported on $[a, b]$ (in practice, given a current estimate of $\boldsymbol{\beta}$, we can set a and b as the minimum and maximum values of $\mathbf{x}_{ij}{}^{\mathrm{T}}\boldsymbol{\beta}$, respectively), we use polynomial splines to approximate the link function. Let $\tau_0 = a < \tau_1 < \cdots < \tau_{k_1'} < b = \tau_{k_1'+1}$ be a partition of $[a, b]$ into subintervals $[\tau_l, \tau_{l+1}), l = 0, \ldots, k_1'$ with k_1' internal knots. We only restrict our attention to equally spaced knots although data-driven choice can be considered such as putting knots at certain sample quantiles of the observed covariate values. A polynomial spline of order s_1 is a function whose restriction to each subinterval is a polynomial of degree $s_1 - 1$ and globally $s_1 - 2$ times continuously differentiable on $[a, b]$. The collection of splines with a fixed sequence of knots forms a B-spline basis $\mathbf{b}_1(x) = \{b_{1,1}(x), \ldots, b_{1,k_1}(x)\}$ with $k_1 = k_1' + s_1$. Similarly, assuming, without loss of generality, that t_{ij} is supported on $[0, 1]$, we can construct B-spline basis $\mathbf{b}_2(t) = \{b_{2,1}(t), \ldots, b_{2,k_2}(t)\}$. We assume B-spline basis is normalized such that $\sum_{l=1}^{k_1} b_{1,l}(x) = \sqrt{k_1}$ and $\sum_{l=1}^{k_2} b_{2,l}(x) = \sqrt{k_2}$. Such normalization is unessential and is only imposed to simplify certain expressions in theoretical derivations later. Finally, for a bivariate function supported on $[a, b] \times [0, 1]$, we construct the tensor basis $\mathbf{b}(x, t) = (b_1(x, t), \ldots, b_k(x, t)) = (b_{1,1}(x)b_{2,1}(t), \ldots, b_{1,k_1}(x)b_{2,k_2}(t))$, where $k = k_1 k_2$.

Using spline approximation, writing $g \approx \mathbf{b}^{\mathrm{T}}\boldsymbol{\theta}$, we minimize

$$l\left(\boldsymbol{\beta}^{(-1)}, \boldsymbol{\theta}\right) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \rho_\tau(y_{ij} - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_{ij}^T\boldsymbol{\beta}, t_{ij})\boldsymbol{\theta}), \tag{1}$$

over $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ with the constraint $\|\boldsymbol{\beta}\| = 1$, or equivalently regard $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\beta}^{(-1)})$ as a function of $\boldsymbol{\beta}^{(-1)}$ and optimize over $(\boldsymbol{\beta}^{(-1)}, \boldsymbol{\theta})$.

An important property regarding splines used in the proof is that the matrix $\{\int b_{2,l}(t)b_{2,l'}(t)dt\}_{l,l'=1}^{k_2}$ possesses eigenvalues bounded and bounded away from zero under our particular normalization for basis functions. Although unnecessary for estimation, setting $b_{2,1}(t) \equiv 1$ is convenient for the testing problem considered in Section 3. Note that since originally we