COMSTA: 6678 Model 3G pp. 1–19 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Pursuit of dynamic structure in quantile additive models with longitudinal data

Xia Cui [a], Weihua Zhao [b,*], Heng Lian [c,*], Hua Liang [d]

[a] School of Economics and Statistics, Guangzhou University, Guangzhou, China
[b] School of Sciences, Nantong University, Nantong, China
[c] Department of Mathematics, The City University of Hong Kong, Kowloon Tong, Hong Kong
[d] Department of Statistics, George Washington University, Washington, DC, USA

ARTICLE INFO

Article history:
Received 1 January 2018
Received in revised form 29 June 2018
Accepted 20 August 2018
Available online xxxx

Keywords:
B-splines
Dynamic structure pursuit
Quantile regression
Sparse functional data

ABSTRACT

We consider quantile additive models with dynamic (time-varying) component functions. We allow some of the component functions to be non-dynamic, and show, as expected but technically nontrivially, that estimators of the non-dynamic functions have a faster convergence rate. A penalization-based method, called dynamic structure pursuit, is proposed to automatically identify these non-dynamic functions. Finally, in the sparse setting, a four-stage estimation procedure is proposed which first identifies the nonzero component functions and then applies the identification strategy of the non-dynamic functions. Theoretical and numerical results are provided to illustrate the performance of the estimators.

1. Introduction

Additive models (AM) provide an efficient way of coping with nonparametric estimation problems and avoid the "curse of dimensionality" in nonparametric problems by assuming that the effects of the different predictors act on the response additively, although possibly nonlinearly. More specifically, with Y the response, $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ the predictors, one assumes

$$Y = \mu + \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

where μ is the intercept parameter, f_j are unknown component functions, and ϵ represents the mean zero noise.

In the early literature, kernel-based backfitting and local scoring procedures have been proposed by Buja et al. (1989) to iteratively estimate the nonparametric components by solving a large system of equations (Yu et al., 2008). Linton and Nielsen (1995) applied the marginal integration approach (Linton and Härdle, 1996) to estimate the parametric components by treating the summand of additive terms as a nonparametric component, which is then estimated as a multivariate nonparametric function. As it is well-known that the kernel-based backfitting and marginal integration approaches are computationally expensive, Wood (2004), Ruppert et al. (2003) and Marx and Eilers (1998) suggested penalized regression splines, which share most of the practical benefits of smoothing spline methods combined with ease of use and reduction of the computational cost of backfitting generalized additive models (GAMs). But no theoretical justifications are available for these procedures.

* Corresponding authors.
E-mail addresses: zhaowhstat@163.com (W. Zhao), henglian@cityu.edu.hk (H. Lian).

https://doi.org/10.1016/j.csda.2018.08.017
0167-9473/© 2018 Elsevier B.V. All rights reserved.

To overcome these limitations, Wang et al. (2011) proposed to estimate the nonparametric components by polynomial splines (Stone, 1986, 1994; Huang, 1998; Xue and Yang, 2006; Andrews, 1991; Andrews and Whang, 1990; Chen, 2007; Donald and Newey, 1994; Newey, 1997; Wang and Tian, 2016; Zhao et al., 2018). After the spline basis is chosen, the coefficients can be estimated by an efficient one-step procedure of maximizing the quasi-likelihood function. Thus the gain of the proposed in terms of computational reduction is remarkable in contrast to alternative estimation methods. In addition, the proposed procedure can easily formulate a penalized functional to implement variable selection. See Wang et al. (2011) for more details.

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, it can uncover different structural relationships between covariates and responses at the upper or lower tails, which is often of significant interest in econometrics and biomedical applications. This inspired some works on quantile additive models recently (Horowitz and Lee, 2005; Lian, 2012; Kato, 2011).

Here we consider functional/longitudinal data model, allowing both responses and predictors to be functional. Zhang et al. (2013) proposed the following dynamic time-varying model:

$$Y(t) = \mu(t) + \sum_{j=1}^{p} f_j(X_j(t), t) + \epsilon(t). \tag{1}$$

The model above is a natural extension of the conventional additive model by allowing time-dynamic bivariate component functions. At any given time t, this reduces to the conventional additive model. The model achieves dimension reduction while being able to capture potential dynamic relations of functional/longitudinal predictors and responses. Zhang et al. (2013) considered the case that the stochastic processes are observed at sparse discrete time points, as we also assume in this work.

The main contributions of our work are four-fold, extending the approach of Zhang et al. (2013) in various ways. First, we consider quantile estimation of (1), which provides a more complete characterization of the conditional distribution of the response. Second, we consider the case that only some, but not all of the component functions are time-dynamic, resulting in a partially dynamic model, and establish (as expected but technically nontrivial) that the estimator for the non-dynamic component functions converges at a faster rate than the dynamic component functions. As far as we know, this is the first paper that established such convergence rate results in models involving both bivariate and univariate functions. Third, we propose a penalization-based framework, called dynamic structure pursuit, for automatically separating the dynamic and non-dynamic component functions. Fourth, under the setting assuming the true model is sparse with some irrelevant predictors, we propose a four-stage penalized estimation procedure that first selects the relevant predictors followed by the dynamic structure pursuit and establish its nonparametric oracle property (Storlie et al., 2011).

The rest of the article is organized as follows. In Section 2, we consider splines-based quantile regression for partially dynamic additive models, assuming the identities of the non-dynamic component functions are known. In the case that the identities of the non-dynamic component functions are unknown, it can be regarded as the infeasible oracle model. For the latter case, in Section 3, we present a penalized estimation method for dynamic structure pursuit. Oracle properties are established which show that the correct structure can be identified with probability approaching one. In Section 4, under the paradigm of sparse modelling, we use a four-stage procedure to separate the zero components, nonzero dynamic components and the nonzero non-dynamic components. Section 5 contains simulation studies and a real data analysis. We conclude in Section 6 with discussions. Finally, The technical proofs are relegated to the Appendix.

2. Estimation of partially dynamic additive models

Under the functional framework, we assume $Y_i(t)$, $\mathbf{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^{\mathrm{T}}$ are i.i.d. stochastic processes across i and

$$Q_{Y_i(t)|X_i(t)}(\tau) = \mu(t) + \sum_{l=1}^{p} f_l(X_{il}(t), t), \tag{2}$$

where $Q_{Y_i(t)|X_i(t)}(\tau)$ denotes the τ-conditional quantile of $Y_i(t)$ conditional on $\mathbf{X}_i(t)$ for $\tau \in (0, 1)$. The dependence of the right-hand side of (2) on τ is suppressed when no confusion arises. For subject i, we observe time points $t_{ij}, j = 1, \ldots, m_i$ and the response and the predictor processes are observed on these time points. Writing $y_{ij} = Y_i(t_{ij})$ and $\mathbf{x}_{ij} = \mathbf{X}_i(t_{ij}) = (x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$, we have

$$Q_{y_{ij}|\mathbf{x}_{ij},t_{ij}}(\tau) = \mu(t_{ij}) + \sum_{l=1}^{p} f_l(x_{ijl}, t_{ij}). \tag{3}$$

There is actually almost no modifications necessary in methodology and theory when we consider longitudinal data, for which we do not regard the response and predictors as stochastic processes, but that we have m_i observations of subject i, $1 \le i \le n$ at time points $t_{ij}, j = 1, \ldots, m_i$. Thus we can still write the quantile regression model as in (3). Under our assumptions stated later, the asymptotic properties for both cases are established in exactly the same way. Without loss

Please cite this article in press as: Cui X., et al., Pursuit of dynamic structure in quantile additive models with longitudinal data. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.017.