COMSTA: 5820 Model 3Gsc pp. 1–18 (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

Partially linear modeling of conditional quantiles using penalized splines[☆]

Q1 Chaojiang Wu[a,*], Yan Yu[b]

[a] Decision Sciences Department, Drexel University, Philadelphia, PA 19104, USA

[b] University of Cincinnati, PO BOX 210130, Cincinnati, OH, 45221, USA

ARTICLE INFO

Article history:
Received 4 April 2013
Received in revised form 24 February 2014
Accepted 26 February 2014
Available online xxxx

Keywords:
Additive model
Dimension reduction
P-splines
Single-index model
Semiparametric model
Smoothing parameter

ABSTRACT

We consider the estimation problem of conditional quantile when multi-dimensional covariates are involved. To overcome the "curse of dimensionality" yet retain model flexibility, we propose two partially linear models for conditional quantiles: partially linear single-index models (QPLSIM) and partially linear additive models (QPLAM). The unknown univariate functions are estimated by penalized splines. An approximate iteratively reweighted penalized least square algorithm is developed. To facilitate model comparisons, we develop effective model degrees of freedom for penalized spline conditional quantiles. Two smoothing parameter selection criteria, Generalized Approximate Cross-validation (GACV) and Schwartz-type Information Criterion (SIC) are studied. Some asymptotic properties are established. Finite sample properties are investigated through simulation studies. Application to the Boston Housing data demonstrates the success of the proposed approach. Both simulations and real applications show encouraging results of the proposed estimators.

1. Introduction

This study is motivated by analyzing the well-known Boston Housing data. The dataset contains 506 observations of the median price of owner-occupied homes together with 13 variables which may affect the median housing prices. The response variable and many covariates are left-skewed. Hence, quantile regression or conditional quantile (see the seminal work Koenker and Bassett, 1978) can be naturally used to examine this data (e.g., Chaudhuri 1997; De Gooijer and Zerom, 2003; Yu and Lu, 2004; Wu et al., 2010; and Kong and Xia, 2012). Q2

Let Y be the response variable and $(\mathbf{X}, \mathbf{Z})$ be the covariate vector. This paper proposes two alternative partially linear models for conditional quantiles: partially linear single-index models (QPLSIM) and partially linear additive models (QPLAM). Both approaches attempt to overcome the well-known "curse of dimensionality" of fully nonparametric model (Chaudhuri, 1991a,b) yet retain model flexibility. Q3

[☆] This manuscript is based in part on the Ph.D. Thesis of the first author at the University of Cincinnati. Part of this paper has been formerly circulated under the title "Partially Linear Single-Index Conditional Quantiles", which was presented in Joint Statistical Meeting (JSM), Miami Beach, FL, August, 2011; INFORMS Annual Meeting, Charlotte, NC, November, 2011; Decision Science Institute Annual Meeting, Boston, MA, November, 2011; and the Eighth ICSA International Conference: Frontiers of Interdisciplinary and Methodological Statistical Research, Guangzhou, December, 2010.

* Corresponding author. Tel.: +1 215 571 4590.
E-mail addresses: chaojiang.wu@gmail.com, Chaojiang.Wu@drexel.edu, wuca@mail.uc.edu (C. Wu), Yan.Yu@uc.edu (Y. Yu).

http://dx.doi.org/10.1016/j.csda.2014.02.020
0167-9473/

The partially linear single-index models for conditional quantiles take the form

$$q_\tau(Y|\mathbf{X}, \mathbf{Z}) = g(\mathbf{X}\boldsymbol{\alpha}) + \mathbf{Z}\boldsymbol{\beta}. \quad (1)$$

QPLSIM has two components: the linear combination $\mathbf{X}\boldsymbol{\alpha}$ which is often termed as single-index enters the model via a univariate nonparametric link function $g(\cdot)$ and $\mathbf{Z}\boldsymbol{\beta}$ enters the model as a partially linear term. By reducing the dimensionality from that of a general covariate vector $\mathbf{X}$ to a univariate single-index $\mathbf{X}\boldsymbol{\alpha}$, QPLSIM avoids the so-called "curse of dimensionality". Partially linear term also enjoys easier interpretation of the effect of each variable. It is worth noting that when the univariate link function $g(\cdot)$ is monotonic, the single-index coefficient $\boldsymbol{\alpha}$ of model (1) retains similar ease-of-interpretation as that in the linear quantile regression while allowing for curvature through the univariate nonparametric link function.

Model (1) is quite flexible to include single-index models and partially linear models as special cases. When there is no partially linear term $\mathbf{Z}\boldsymbol{\beta}$, model (1) reduces to the single-index quantile models (Wu et al., 2010). Based on their estimation, Kong and Xia (2012) investigated the Bahadur representation of single-index parameter estimators. Both papers adopt local linear methods to estimate the univariate link function $g(\cdot)$. Chaudhuri et al. (1997) proposed the average derivative approach, where the single-index coefficient is estimated by taking an expectation of the vector of partial derivatives of the conditional quantile with respect to the covariates. Since this approach requires multivariate kernel regression, it is less appealing in practice with high-dimensional covariates. In fact, when the single-index coefficient is known or the covariate $\mathbf{X}$ is univariate, model (1) reduces to partially linear models. For example, He and Liang (2000) studied partially linear errors-in-variables quantile models. Chen and Khan (2001) studied partially linear censored quantile regression. Lee (2003) studied the efficient estimator for partially linear terms. Most recently, Härdle et al. (2012) studied the bootstrap approximation for the uniform confidence band.

The second approach we propose is partially linear additive models for quantile regression (QPLAM)

$$q_\tau(Y|\mathbf{X}, \mathbf{Z}) = a + g_1(X_1) + \cdots + g_d(X_d) + \mathbf{Z}\boldsymbol{\beta}, \quad (2)$$

which include additive models and partially linear models as special cases. Additive models are popular for dimension reduction which replace the linear component by a sum of univariate nonparametric functions over the components of $\mathbf{X}$; see e.g. De Gooijer and Zerom (2003) on kernel based marginal integration, Yu and Lu (2004) on local linear smoothing, Horowitz and Lee (2005) on two-stage estimator for quantile regression.

Note that, QPLAM model (2) reduces dimensionality nicely but does not incorporate interactions of $\mathbf{X}$ in the current structure. QPLSIM model (1) applies a nonlinear link function to the index $\mathbf{X}\boldsymbol{\alpha}$, hence some interactions between the covariates can be modeled. Computationally QPLSIM are more complex partly due to this nonlinear structure on index parameters where nonlinear optimization has to be involved. In this study, we find that both approaches are appealing alternatives for dimension reduction on conditional quantile regression.

We estimate the nonparametric functions in models (1) and (2) using penalized splines (P-splines). Penalized splines have gained increasing popularity, especially due to the computational expediency and easy-adaptability to more complex models (see Ruppert et al., 2003 for a review). Yu and Ruppert (2002) show that penalized spline estimation for partially linear single-index models in mean regression is computationally stable and expedient, while the local method estimation may become computationally unstable (Carroll et al., 1997). To the best of our knowledge, this is the first work to adopt the penalized spline estimation for single-index quantile regression.

Furthermore, we study the effective degrees of freedom in P-splines quantile regression models and investigate the smoothing parameter selection in detail. First, to facilitate model comparisons, we develop the measure for effective model degrees of freedom for the QPLSIM and QPLAM via an approximate iteratively reweighted least square algorithm. The main idea of our algorithm is centered on minimizing a check function along with a roughness penalty for models (1) and (2), where the check function is approximated by a differentiable function near a small neighborhood of the origin. This idea can be traced back to Nychka et al. (1995) and has also been adopted in Yuan (2006) and Li et al. (2007) in different modeling contexts. Second, we study in simulations two smoothing parameter selection criteria, Generalized Approximate Cross-validation (GACV) and Schwarz-type information criterion (SIC).

Incorporating model degrees of freedom, we compare model goodness-of-fit for a variety of conditional quantile models with Boston Housing data. The findings are interesting. Among the single-index models, QPLSIM models outperform single-index models without partially linear terms. Similar results are observed in the additive models. Both observations suggest that incorporating model degrees of freedom, partially linear terms are useful to include in modeling conditional quantiles. Among the two partially linear models, QPLSIM and QPLAM are comparable in terms of model goodness-of-fit (see Tables 6 and 7). Finally, simulation studies with various error distributions show P-splines estimator for QPLSIM performs well. Simulations with identical design as De Gooijer and Zerom (2003) and Horowitz and Lee (2005) show the superiority of P-splines estimator to the existing additive quantile estimators (see Table 3).

The rest of the paper is organized as follows. Section 2 describes the P-spline estimation for the partially linear single-index models (QPLSIM), establishes some asymptotic properties of the proposed estimators and presents the simulation results. Section 3 provides the study of the partially linear additive conditional quantile models (QPLAM). Section 4 presents an application to the Boston Housing data. Section 5 concludes the paper and discusses the link of our work to the research area of functional data analysis. Technical proofs are relegated to the Appendix.