Variable Selection for Nonparametric Quantile Regression via Smoothing Spline ANOVA

Chen-Yen Lin, Hao Helen Zhang, Howard D. Bondell and Hui Zou

February 15, 2012

Author's Footnote:

Chen-Yen Lin (E-mail: clin5@ncsu.edu) is currently a PhD student, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203. Dr. Hao Helen Zhang (E-mail: hzhang2@stat.ncsu.edu) and Dr. Howard D. Bondell (E-mail: bondell@stat.ncsu.edu) are Associate professors, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203. Dr. Hui Zou (E-mail: hzou@stat.umn.edu) is Associate professor, Department of Statistics, University of Minnesota, Minneapolis MN, 55455

Abstract

Quantile regression provides a more thorough view of the affect of covariates on a response. In many cases, assuming a parametric form for the conditional quantile can be overly restrictive. Nonparametric quantile regression has recently become a viable alternative. The problem of variable selection for quantile regression is challenging, since important variables can influence various quantiles in different ways. We propose to tackle the problem using the approach of nonparametric quantile regression via regularization in the context of smoothing spline ANOVA models. By imposing the sum of the reproducing kernel Hilbert space norms on functions, the proposed sparse nonparametric quantile regression (SNQR) can identify variables which are important in either conditional mean or conditional variance, and provide flexible nonparametric estimates for quantiles. We develop an efficient algorithm to solve the optimization problem and contribute an R package. Our numerical study suggests the promising performance of the new procedure in variable selection for heteroscedastic data analysis.

KEYWORDS: Quantile Regression, Smoothing Spline ANOVA, Model Selection

1. INTRODUCTION

Quantile regression, as a complement to classical least square regression, provides a more comprehensive framework to study how covariates influence not only the location but the entire conditional distribution (Koenker 2005). In quantile regression problems, the primary interest is to establish a regression function to reveal how the $100\tau\%$ quantile of the response Y depends on a set of covariates $\boldsymbol{X} = \{X^{(1)}, \ldots, X^{(d)}\}$. A parametric form of regression function is often assumed for convenience of interpretation and lower computational cost. While a linear regression function is studied in Koenker and Bassett (1978) and numerous follow-up studies, Procházka (1988) and Jurečková and Procházka (1994) explored nonlinear regression; see Koenker and Hallock (2001) and Koenker (2005) for a comprehensive overview.

As much as the parametric assumption enjoys a simple model structure and lower cost for implementation, it is not flexible enough and hence carries the risk of model misspecifications for complex problems. For a single predictor model, Koenker, Ng and Portnoy (1994) pioneered nonparametric quantile regression in spline models, in which the quantile function can be found via solving the minimization problem

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau(y_i - f(x_i)) + \lambda V(f'), \tag{1}$$

where $\rho_\tau(\cdot)$ is the so-called "check function" of Koenker and Bassett (1978),

$$\rho_\tau(t) = t[\tau - I(t < 0)], \;\; \tau \in (0, 1) \tag{2}$$

λ is a smoothing parameter and $V(f')$ is the total variation of the derivative of f. Koenker et al. (1994) showed that the minimizer is a linear spline with knots at the design points $x_i, i = 1, \ldots, n$ provided that the space $\mathcal{F}$ is an expanded second-order Sobolev space defined as

$$\mathcal{F} = \left\{ f : f(x) = a_0 + a_1 x + \int_0^1 (x - y)_+ d\mu(y), \;\; V(\mu) < \infty, a_i \in \mathbb{R}, i = 0, 1 \right\} \tag{3}$$

where μ is a measure with finite total variation.

For multi-dimensional feature space, He, Ng and Portnoy (1998) proposed a bivariate quantile smoothing spline and He and Ng (1999) generalized the idea to multiple covariates using an ANOVA-type decomposition. More recently, Li et al. (2007) proposed a more general framework called the kernel quantile regression (KQR). By penalizing the roughness of the function estimator using its squared functional norm in a reproducing kernel Hilbert space (RKHS), the KQR solves the regularization problem

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} \rho_\tau(y_i - f(\boldsymbol{x}_i)) + \frac{\lambda}{2} ||f||^2_{\mathcal{H}_K} \tag{4}$$

where $\mathcal{H}_K$ is a RKHS and $|| \cdot ||_{\mathcal{H}_K}$ is the corresponding function norm.

Despite several existing nonparametric quantile function estimators, selecting relevant predictors in multi-dimensional data is an important yet challenging topic that has not been addressed in depth. Variable selection in quantile regression is much more difficult than that in the least square regression. When the error distribution is not symmetric, the variable selection is carried at various levels of quantiles, which amounts to identifying variables that are important for the entire distribution, rather than limited to the mean function as in the least squares regression case. This has important applications to handle heteroscedastic data. Several regularization methods were proposed (Zou and Yuan 2008*a*; Zou and Yuan 2008*b*; Wu and Liu 2009) for linear quantile regression. However, to our knowledge, there still lacks of study on variable selection in nonparametric quantile regression. This is the main motivation of our work.

In the presence of multiple predictors, many nonparametric estimation procedures may suffer from the curse of dimensionality. The smoothing spline analysis of variance (SS-ANOVA) models (Wahba 1990) provide a flexible and effective estimation framework to tackle the problem. Since some of the predictors may not be useful or redundant for prediction, variable selection is important in nonparametric regression. In the context of least squares regression, the COmponent Selection and Shrinkage Operator (COSSO, Lin and Zhang (2006)) was proposed to perform continuous function shrinkage and estimation by penalizing the sum of RKHS norms of the components. However, variable selection in nonparametric quantile regression is void in the literature. In this paper, we adopt the COSSO-type penalty to develop a new penalized framework for joint quantile estimation and variable selection. An adaptive penalty is then considered to further enhance the solution sparsity.

The remainder of the article is organized as follows. Section 2 reviews the SS-ANOVA models and introduces the new estimator. An iterative computation algorithm is given in Section 3, along with parameter tuning procedure. Extensive empirical studies, including both the homogeneous and heterogenous errors are given in Section 4. Three real example analysis results are presented in Section 5. We conclude our findings in Section 6.

2. FORMULATION

2.1 Smoothing Spline ANOVA

In the framework of smoothing spline ANOVA (SS-ANOVA), it is assumed that a function $f(\boldsymbol{x}) = f(x^{(1)}, \ldots, x^{(d)})$ has the ANOVA decomposition

$$f(\boldsymbol{x}) = b + \sum_{j=1}^{d} f_j(x^{(j)}) + \sum_{j<k} f_{j,k}(x^{(j)}, x^{(k)}) + \cdots \tag{5}$$

where b is a constant, f_j's are the main effects and $f_{j,k}$'s are the two-way interactions, and so on. We estimate each of the main effects in a RKHS denoted by $\mathcal{H}_j = \{1\} \oplus \bar{\mathcal{H}}_j$ whereas the interactions are estimated in a tensor product spaces of the corresponding univariate function spaces. A popular choice of $\mathcal{H}_j$ is the second-order Sobolev space $\mathcal{S}^2[0,1] = \{g : g, g' \text{ are absolutely continuous and } g'' \in \mathcal{L}^2[0,1]\}$. When endowed with the norm

$$||f||^2 = \left\{ \int_0^1 f(x)dx \right\}^2 + \left\{ \int_0^1 f'(x)dx \right\}^2 + \int_0^1 \{f''(x)\}^2 dx \tag{6}$$

the second order Sobolev space is a RKHS with reproducing kernel

$$R(x,y) = 1 + k_1(x)k_1(y) + k_2(x)k_2(y) - k_4(|x-y|) \tag{7}$$

where $k_1(x) = x - \frac{1}{2}$, $k_2(x) = \frac{1}{2}\left[k_1^4(x) - \frac{1}{12}\right]$ and $k_4(x) = \frac{1}{24}\left[k_1^4(x) - \frac{1}{2}k_1^2(x) + \frac{7}{240}\right]$. See (Wahba 1990; Gu 2002) for more details. The entire tensor-product space for estimating $f(\boldsymbol{x})$ is given by

$$\mathcal{F} = \otimes_{j=1}^{d} \mathcal{H}_j = \{1\} \oplus \sum_{j=1}^{d} \bar{\mathcal{H}}_j \oplus \sum_{j<k} \left(\bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k\right) \oplus \cdots \tag{8}$$

Note that $\mathcal{F} = \otimes_{j=1}^{d} \mathcal{H}_j$ is also a RKHS, and its reproducing kernel is the sum of the reproducing kernels of those component spaces.

In practice, the higher-order interactions in (4) will usually be truncated for convenience in interpretation and to avoid the curse of dimensionality. A general expression for a truncated space can be written as

$$\mathcal{F} = \{1\} \oplus \mathcal{F}_1 = \{1\} \oplus \left\{\oplus_{j=1}^{q} \mathcal{F}_j\right\} \tag{9}$$

where $\mathcal{F}_1, \ldots, \mathcal{F}_q$ are q orthogonal subspaces of $\mathcal{F}$. A special case is the well-known additive model (Hastie and Tibshirani 1990) with $q = d$, in which only the main effects are kept in the model, say $f(\boldsymbol{x}) = b + \sum_{j=1}^{d} f_j(x^{(j)})$. When both main effects and two-way interaction effects are retained, the truncated space has $q = d(d+1)/2$. For illustration purpose, we focus on the main effects model in this paper. The proposed idea can be naturally generalized to any function space with higher order interactions.

A typical method for estimating nonparametric quantile function is through solving the regularization problem

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - f(\boldsymbol{x}_i)) + \lambda J(f) \tag{10}$$

where λ is a smoothing parameter and $J(\cdot)$ is a penalty functional. A smoothing spline estimate uses the penalty $J(f) = \sum_{j=1}^{d} \theta_j^{-1} ||P^j f||^2$, with θ_j's as smoothing parameters. The estimation in (10) involves multiple tuning parameters $\theta_1, \cdots, \theta_d$, which needs to be selected properly for a good estimation results. The parameter λ is usually included and fixed at some convenient value for computational stability in practice.

2.2 New Methodology: Sparse Nonparametric Quantile Regression (SNQR)

To achieve joint variable selection and function estimation in nonparametric quantile regression, we consider the following regularization problem

$$\frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - f(x_i)) + \lambda \sum_{j=1}^{d} w_j ||P^j f||_{\mathcal{F}}, \tag{11}$$

where the penalty functional penalizes the sum of component norms. The problem (11) reduces to the L_1-norm quantile regression (Li and Zhu 2008) in the special case of linear models. In particular, if $f(\boldsymbol{x}) = b + \sum_{j=1}^{d} \beta_j x^{(j)}$ and we consider a linear function space $\mathcal{F} = \{1\} \oplus \left\{\oplus_{j=1}^{d}\{x^{(j)} - 1/2\}\right\}$ with inner product $\langle f, g\rangle = \int fg$, then the RKHS norm penalty $||P^j f||^2$ becomes proportional to $|\beta_j|$. We allow each functional component to be penalized differently depending on its associated weight $w_j \in (0, \infty)$. In principle, smaller weights are assigned to important function components while larger weights are assigned to less important components. This is in the same spirit of the adaptive LASSO (Zou 2006) and adaptive COSSO (Storlie, Bondell, Reich and Zhang 2011). We propose to construct the weights w_j from the data adaptively, For each component $f^{(j)}$, its L_2 norm $||f^{(j)}(x)||_{L_2} = \sqrt{\int [f^{(j)}(x)]^2 dF(x)}$ is a natural measure to quantify the importance of functional components. In practice, given a reasonable initial estimator $\tilde{f}$, we propose to construct the weights w's by

$$w_j = ||P^j \tilde{f}||_{n,L_2} = \sqrt{n^{-1} \sum_{i=1}^{n} [P^j \tilde{f}(\boldsymbol{x}_i)]^2}, \quad \forall j. \tag{12}$$

A convenient choice of $\tilde{f}$ is the solution of the KQR.

Due to the fact that both the check loss and the penalty functional $J(f)$ are continuous and convex in f, the existence of the minimizer of (11) is guaranteed as stated in the following Theorem.

Theorem 1. *Let $\mathcal{F}$ be an RKHS of functions with the decomposition (9), then there exists a minimizer to (11) in $\mathcal{F}$.*

Directly minimizing (11) can be a daunting task as searching over the infinite dimensional space $\mathcal{F}$ for a minimizer is practically infeasible. Analogous to the smoothing spline models, the following theorem shows that the minimizer to (11) lies in a finite dimensional space. This important result assures the feasibility of computation.

Theorem 2. *Representer Theorem: Let the minimizer of (11) be $\hat{f} = \hat{b} + \sum_{j=1}^{d} \hat{f}_j$ with $\hat{f}_j \in \bar{\mathcal{H}}_j$, then $\hat{f}_j \in span\{R_j(x_i, \cdot), i = 1, \ldots, n\}$ where $R_j(\cdot, \cdot)$ is the reproducing kernel of $\mathcal{F}_j$.*

3. ALGORITHM

To further facilitate the computation, we first present an equivalent formulation of (11). By introducing non-negative slack variables $\theta_j, j = 1, \ldots, d$, it is easy to show that minimizing (11) is equivalent to solving the following optimization problem

$$\min_{f, \boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - f(x_i)) + \lambda_0 \sum_{j=1}^{d} w_j^2 \theta_j^{-1} ||P^j f||^2 \text{ s.t. } \sum_{j=1}^{d} \theta_j \leq M, \tag{13}$$

where λ_0 and M are both smoothing parameters. As opposed to smoothing spline models, the non-negative θ_j's shed the light on recovering the sparse structure. Based on Theorem 2, we can show that the minimizer to (11) has the expression

$$\hat{f}(\boldsymbol{x}) = \hat{b} + \sum_{i=1}^{n} \hat{c}_i \sum_{j=1}^{d} \frac{\hat{\theta}_j}{w_j^2} R_j(\boldsymbol{x}_i, \boldsymbol{x}) \tag{14}$$

where $\hat{\boldsymbol{c}} = (\hat{c}_1, \ldots, \hat{c}_n) \in \mathbb{R}^n$ and $\hat{b} \in \mathbb{R}$. Clearly, if $\hat{\theta}_j = 0$, the minimizer does not depend on the j-th reproducing kernel, implying the function component f_j or variable X_j being irrelevant.

Define $\boldsymbol{R}^\theta = \sum_{j=1}^{d} w_j^{-2} \theta_j R_j(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$, the $n \times n$ matrix with elements $R_{i,i'}^\theta$. Plugging (14) into (13), the objective function becomes

$$\min_{b, \boldsymbol{c}, \boldsymbol{\theta}} \ \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(y_i - b - \sum_{k=1}^{n} c_k R_{ki}^\theta \right) + \lambda_0 \boldsymbol{c}^T \boldsymbol{R}^\theta \boldsymbol{c} \text{ s.t. } \sum_{j=1}^{d} \theta_j \leq M. \tag{15}$$

For the remaining of the article, we will refer to (15) as the objective function of our proposed method.

3.1 Iterative Optimization Algorithm

It is possible to minimize the objective function in (15) with respect to all the parameters, $(b, \boldsymbol{c}, \boldsymbol{\theta})$, simultaneously, but the programming effort can be substantial. Alternatively, we can decompose the parameters into parts $\boldsymbol{\theta}$ and $(b, \boldsymbol{c}^T)^T$ and iteratively solve two sets of optimization problems in turn, with respect to $\boldsymbol{\theta}$ and $(b, \boldsymbol{c}^T)^T$. Consequently, we suggest the following iterative algorithm:

1. Fix $\boldsymbol{\theta}$, solve $(b, \boldsymbol{c}^T)^T$

$$\min_{b,\boldsymbol{c}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(y_i - b - \sum_{k=1}^{n} c_k R_{ki}^{\theta} \right) + \lambda_0 \boldsymbol{c}^T \boldsymbol{R}^{\theta} \boldsymbol{c} \tag{16}$$

2. Fix $(b, \boldsymbol{c}^T)^T$,solve

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(y_i^* - \sum_{j=1}^{d} \theta_j G_{ij} \right) + \lambda_0 \boldsymbol{c}^T \boldsymbol{G} \boldsymbol{\theta}, \text{ s.t. } \sum_{j=1}^{d} \theta_j \leq M, \theta_j \geq 0 \tag{17}$$

   where $y_i^* = y_i - b$ and G_{ij} is an element from the matrix $\boldsymbol{G} = \left(w_1^{-2} R_1 \boldsymbol{c}, \ldots, w_d^{-2} R_d \boldsymbol{c} \right)_{n \times d}$.

The optimization problems in (16) and (17) can be cast into quadratic programming and linear programming problems, respectively. We defer all the derivations to the Appendix. So, both of them can be solved using standard optimization softwares, such as MATLAB and `R`.

In practice, based on our empirical experience, the algorithm converges quickly in a few steps. We have noted that the one-step solution often provides a satisfactory approximation to the solution. As a result, we advocate the use of one-step update in practice.

An important connection between our proposed method and the KQR can be unraveled by realizing that the objective function in (16) is exactly the same as that in the KQR. This connection suggests that when $\boldsymbol{\theta}$ is known, our proposed method shares the same spirit as the KQR. The optimization problem for estimating $\boldsymbol{\theta}$ essentially imposes the non-negative garrote (Breiman 1995) type shrinkage on θ's, and hence achieves variable selection by shrinking some of θ_j's to zero.

3.2 Parameter Tuning

Like any other penalized regression problem, the performance of the new estimator critically depends on properly-tuned smoothing parameters in (15). Smoothing parameters play an important role in balancing the trade-off between the goodness of data fit and the model complexity. A reasonable parameter choice is usually the one that minimizes some generalized error or information criterion. In the quantile regression literature, one commonly used criterion is the Schwarz information criterion (SIC) (Schwarz 1978; Koenker et al. 1994)

$$\log\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{f}(\boldsymbol{x}_i))\right) + \frac{\log n}{2n}df \tag{18}$$

where df is a measure of complexity of the fitted model. Koenker et al. (1994) and Li, Liu and Zhu (2007) argued using the number of interpolated y_i's as an estimate of effective degree of freedom.

In addition to the SIC, another popular criterion to choose the smoothing parameter is k-fold cross validation, which has been widely applied to various regression and classification problems and usually gives competitive performance. After numerous experiments, we have observed that the tuning with k-CV overall produces better prediction performance than SIC for the proposed nonparametric quantile estimation. Therefore, we suggest to use k-fold cross validation in practice, where $k = 5$ is used in both simulated and real data analysis.

In the following, we summarize the complete algorithm for the proposed method, including both model fitting and parameter tuning steps.

Step 1. Initialization. Set $\theta_j = 1, \ \forall j$.

Step 2. For each of the grid points of λ_0, solve (16) for $(b, \boldsymbol{c}^T)^T$, and record the SIC score or CV error. Choose the best λ_0 that minimizes the SIC or CV error, then fix it in later steps.

Step 3. For each of the grid points of M, solve (15) for $(\boldsymbol{\theta}^T, b, \boldsymbol{c}^T)^T$ using the aforementioned iterative optimization algorithm. Record the SIC score or CV error at each grid point and choose the best M that minimizes either SIC score or CV error.

Step 4. Solve (15) using the chosen λ_0 and M pair, on the full data. Note that this is already done if tuning was based on SIC.

Since the tuning procedure described above does not cover all the possible pairs (λ_0, M), it would be beneficial to enhance the tuning with a refined procedure. In particular, we suggest to do the following. After Step 3, say, we obtain the optimal pair $(\lambda_0^*, M*)$. Then we focus on a narrowed and more focused region, the neighborhood of (λ_0^*, M^*) and apply Step2 and 3 again. The optimal parameters determined at this refined step, say, (λ_0^{**}, M^{**}) will be used as the final selection. Our simulation study also confirms that this refined tuning procedure can improve the prediction and selection performance substantially.

4. NUMERICAL RESULTS

In this section we present the empirical performance of the SNQR procedure using simulated data. For the experiment design, we use the following functions as building blocks: $g_1(t) = t$; $g_2(t) = (2t-1)^2$; $g_3(t) = \frac{\sin(2\pi t)}{2-\sin(2\pi t)}$ and $g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t)$. Similar settings were also considered in Lin and Zhang (2006).

We evaluate the performance from two aspects, prediction accuracy and model selection. The integrated absolute error (IAE), defined as IAE $= E|\hat{f}(\mathbf{X}) - f(\mathbf{X})|$, is used to assess prediction accuracy, where the expectation is evaluated by a monte carlo integration with 10,000 test points generated from the same distribution as the training data. In terms of model selection, we first denote $\hat{\mathcal{M}} = \{j : \hat{\theta}_j \neq 0\}$ and $\mathcal{M}_0 = \{j : ||P^j f|| > 0\}$ as the selected model and true model, respectively, and $|\mathcal{M}|$ as the cardinality of the set $\mathcal{M}$. Then we compute three statistics for assessing selection accuracy: type I error rate, $\frac{|\hat{\mathcal{M}} \cap \mathcal{M}_0^c|}{d - |\mathcal{M}_0|}$, power, $\frac{|\hat{\mathcal{M}} \cap \mathcal{M}_0|}{|\mathcal{M}_0|}$, and model size, $|\hat{\mathcal{M}}|$. For the purpose of comparison, we also include the solution of the KQR fitted with only relevant predictors based on 5-fold cross validation tuning. This method will later be referred to as the Oracle estimator. The oracle estimator provides a benchmark for the best possible estimation risk if the important variables were known. We also include the KQR for a comparison, to illustrate that including many noisy variables could undermine the prediction accuracy of the KQR.

Another property that we would like to study is the role of the adaptive weights in the performance of the SNQR procedure. Without any a priori knowledge on the importance of each predictor, we can set all $w_j = 1$ in (11) and proceed to solve the objective function. This method will be referred to as an unweighted method. For the weighted procedure, we use KQR as an initial estimate, $\tilde{f}$, to produce an adaptive weight.

A training set of size 200 and three different quantile values τ =20%, 50% and 80%, are used throughout the simulation. For each of the following examples, we repeat 100 times and report the average summary statistics and their associated standard errors. We made extensive comparisons between SIC and 5-fold CV in our simulations studies, and have observed that the SIC-based estimator performs similarly but slightly worse than the CV-based estimator in most cases. For simplicity, we show only the results using cross validation as the tuning procedure.

4.1 Homoskedastic Error Example

Let $\boldsymbol{X} = \{X^{(1)}, \dots, X^{(40)}\} \in [0,1]^{40}$ where each $X^{(j)}$ is independently generated from $U(0,1)$. The response is generated from

$$y = 5g_1(x^{(1)}) + 3g_2(x^{(2)}) + 4g_3(x^{(3)}) + 6g_4(x^{(4)}) + \varepsilon \tag{19}$$

where $\varepsilon \sim t(3)$. Thus $X^{(5)}, \dots, X^{(40)}$ are noise variables. Table 1 presents the performance of four procedures: KQR, unweigted, weighted, and oracle procedure, for $\tau = 0.2, 0.5, 0.8$.

From Table 1, in terms of prediction error, the adaptive SNQR has the smallest IAE and is hence the best, the unweighted SNQR is the second best, and the KQR is the worst. It is clear that the KQR suffers considerably from those noisy variables. The performance of the adaptive SNQR is quite close to the oracle procedure, suggesting that the weights can be useful to improve the procedure under this setting.

With regard to variable selection, the proposed SNQR is effective in identifying important variables and removing noisy variables , which is shown by its small Type I error and large power. Overall speaking, the unweighted SNQR and the adaptive SNRQ perform similarly in terms of Type I error and power. The model size of unweighted SQR is close to the true size 4, suggesting its high accuracy in variable selection. The adaptive method tends to include 1 more irrelevant predictors in the model, so it model size is relatively larger than the unweighted method. Overall speaking, the SNQR procedures show promising performance in terms of both variable selection and quantile estimation.

Table 1: Summary of prediction and estimation results for the homoskedastic example. The standard errors are given in the parentheses.

τ	Method	Type I Error	Power	Model Size	IAE
0.2	KQR	-	-	-	2.283 (0.023)
	unweighted SNQR	0.01 (0.01)	0.98 (0.01)	4.22 (0.10)	0.809 (0.020)
	Adaptive SNQR	0.04 (0.01)	0.98 (0.01)	5.50 (0.23)	0.706 (0.020)
	Oracle	-	-	-	0.632 (0.011)
0.5	KQR	-	-	-	1.923 (0.019)
	unweighted SNQR	0.00 (0.00)	1.00 (0.00)	4.14 (0.06)	0.582 (0.012)
	Adaptive SNQR	0.04 (0.01)	1.00 (0.00)	5.34 (0.24)	0.495 (0.013)
	Oracle	-	-	-	0.493 (0.008)
0.8	KQR	-	-	-	2.302 (0.028)
	unweighted SNQR	0.01 (0.00)	0.99 (0.01)	4.19 (0.09)	0.771 (0.021)
	Adaptive SNQR	0.05 (0.01)	1.00 (0.00)	5.82 (0.27)	0.690 (0.017)
	Oracle	-	-	-	0.635 (0.012)

Figure 1 gives a graphical illustration for the fitted curve and pointwise confidence band given by the adaptive SNQR for $\tau = 0.2$. For comparison, the estimated functions by the Oracle are also depicted. We apply each procedure to 100 simulated datasets and a pointwise confidence band is given by the 5% and 95% percentiles.

Figure 1 suggests that the SNQR procedure produces a very good estimation for the true functions, and the fits are comparable to those given by the Oracle estimator. The fourth function component is more difficult to estimate due to its subtle features in extreme values and inflexion points.

4.2 Heteroskedastic Error Model with Low Dimension

We consider an example with heteroskedastic errors, i.e., the distribution of the error term depends on the covariate $\mathbf{x}$. The covariates are simulated the same way as that in previous example except the dimension is reduced to 10, then we generate response Y from the model

$$y = 5g_1(x^{(1)}) + 3g_2(x^{(2)}) + 4g_3(x^{(3)}) + 6g_4(x^{(4)}) + \exp\left[2g_3(x^{(5)})\right]\varepsilon \tag{20}$$

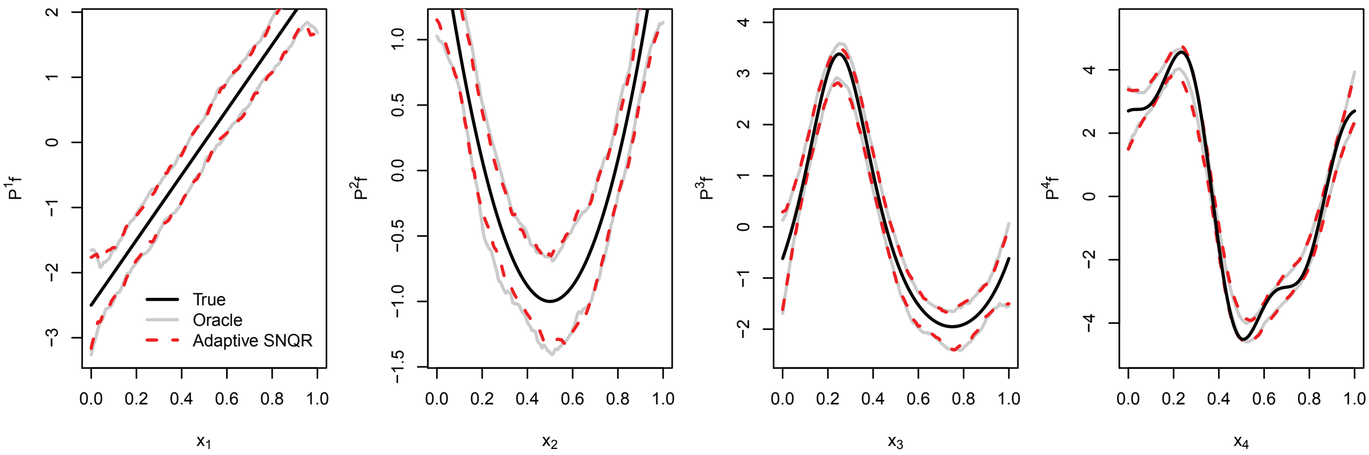

Figure 1: The fitted function components and the associated with pointwise confidence band in Example 1.

where $\varepsilon \sim \mathcal{N}(0,1)$. It follows that the $100\tau\%$ quantile function has the form $5g_1(x^{(1)}) + 3g_2(x^{(2)}) + 4g_3(x^{(3)}) + 6g_4(x^{(4)}) + \exp\left[2g_3(x^{(5)})\right]\Phi^{-1}(\tau)$, where Φ is the distribution function of standard normal. As a result, except for the median, the quantile function will depend on $X^{(5)}$ together with $X^{(1)}$ to $X^{(4)}$. From this example, we aim to evaluate performance of the SNQR for detecting important variables in heteroskedastic data.

Table 2 summarizes the performance of KQR, unweigted SNQR, adaptive SNQR, and the oracle procedure. Again, we observe that the adaptive SNQR is the best in terms of IAE, the unweighted SNQR is the second best, and the KQR is the worst. The prediction accuracy of the adaptive SNQR is overall close to the oracle procedure. In general, the Oracle estimator is better than our methods by 10% to 20%. For variable selection, the SNQR procedures overall have small Type I error and large power. When $\tau = 0.5$, the estimated model size is close to 4 as expected, since the median only depends on the first four predictors. When τ is away from 0.5, both SNQR procedures can successfully identify the additional informative predictor in the error variance, suggesting that the new method's capability to identify all the relevant predictors that influence the distribution of the response.

Table 2: Summary of the prediction and estimation results for the heteroskedastic example. The standard errors are given in the parentheses.

τ	Method	Type I Error	Power	ModelSize	IAE
0.2	KQR	-	-	-	1.268 (0.025)
	Unweighted SNQR	0.04 (0.01)	0.98 (0.01)	5.09 (0.08)	0.936 (0.027)
	Adaptive SNQR	0.02 (0.01)	0.98 (0.01)	5.00 (0.05)	0.872 (0.024)
	Oracle	-	-	-	0.769 (0.018)
0.5	KQR	-	-	-	1.268 (0.011)
	Unweighted SNQR	0.03 (0.01)	1.00 (0.00)	4.16 (0.06)	0.514 (0.011)
	Adaptive SNQR	0.02 (0.01)	1.00 (0.00)	4.08 (0.04)	0.473 (0.010)
	Oracle	-	-	-	0.517 (0.009)
0.8	KQR	-	-	-	1.248 (0.024)
	Unweighted SNQR	0.03 (0.01)	0.97 (0.01)	5.03 (0.10)	0.956 (0.030)
	Adaptive SNQR	0.02 (0.01)	0.97 (0.01)	4.94 (0.11)	0.915 (0.031)
	Oracle	-	-	-	0.861 (0.015)

4.3 Heteroskedastic Error Model in Larger Dimension

To further examine the finite sample performance of the new methods, we consider another heteroskedastic model with a larger dimension of covariates. Combining Example 1 and 2, we generate the response from (20) but $\varepsilon \sim \mathcal{N}(0,1)$.

Table 3: Summary of prediction and estimation results for larger dimensional data with heteroskedastic error. The standard errors are given in the parentheses.

τ	Method	Type I Error	Power	Model Size	IAE
0.2	KQR-5CV	-	-	-	2.995 (0.019)
	Unweighted SNQR	0.02 (0.00)	0.79 (0.01)	4.58 (0.17)	1.673 (0.025)
	Adaptive SNQR	0.06 (0.01)	0.87 (0.01)	6.29 (0.30)	1.552 (0.033)
	Oracle	-	-	-	0.865 (0.016)
0.5	KQR-5CV	-	-	-	2.353 (0.018)
	Unweighted SNQR	0.01 (0.00)	0.96 (0.01)	4.15 (0.11)	0.737 (0.017)
	Adaptive SNQR	0.02 (0.00)	1.00 (0.00)	4.83 (0.15)	0.581 (0.014)
	Oracle	-	-	-	0.494 (0.007)
0.8	KQR-5CV	-	-	-	3.044 (0.022)
	Unweighted SNQR	0.02 (0.00)	0.78 (0.01)	4.45 (0.14)	1.662 (0.024)
	Adaptive SNQR	0.04 (0.01)	0.84 (0.01)	5.74 (0.23)	1.564 (0.030)
	Oracle	-	-	-	0.865 (0.017)

From Table 3, we conclude that the adaptive SNQR provides the best function estimate, the unweighted SNQR is the second best, and the KQR is the worst. The prediction accuracy and model selection are clearly undermined by the enlarged feature space, particularly in both tails. In terms of variable selection, both SNQR procedures miss the additional informative predictor more frequently than before, resulting in reduced power. However, the adaptive weight certainly helps the SNQR procedure so that the adaptive SNQR could better identify the prominent components. When median is of interest, the larger dimension does not affect the performance too much, but these is a greater effect in the tails.

5. REAL DATA ANALYSIS

We apply the SNQR method to three real datasets: prostate cancer data, ozone data and waste isolation pilot plant (WIPP) data. The prostate data is from Stamey et al. (1989), consisting of 97 patients who were about to receive a radical prostatectomy. This data was used by Tibshirani (1996) to model the mean function of the level of prostate-specific antigen on 8 clinical outcomes and select relevant variables. The ozone data contains 330 observations collected in Los Angeles in 1976, and the purpose of the study is to model the relationship between the daily ozone concentration and 8 meteorological covariates. The data has been used in various studies (Buja, Hastie and Tibshirani 1989; Breiman 1995; Lin and Zhang 2006). The WIPP data, with $n = 300$, comes from a computer model for two-phase fluid flow analysis. The response variable, BRNEPTC10K, stands for the cumulative brine flow into the water repository at 10,000 years assuming there is a drilling intrusion at 1000 years. After discarding two highly correlated predictors, we regress the response on 29 predictors; a complete list of predictors and description can be found in Storlie and Helton (2008). The first two data are available in `R` packages `ElemStatLearn` and `cosso`, respectively.

We apply our methods on these datasets and estimate the prediction risk, $E\rho_\tau(Y - f(\boldsymbol{X}))$, by randomly reserving 10% of the data as testing set. The smoothing parameters and model parameters are selected using only the training set. The estimated parameters will then be applied on the testing set and the prediction risk is used as a comparison between various methods. The entire procedure is repeated 100 times and averaged.

Table 4 summarizes the prediction risk along with its associated standard error. Based on the result, the adaptive weights is not always helpful in real application. The advantage of adaptive weight is clear for the WIPP data. However, the unweighted method could predict more accurately in other cases. Nonetheless, the difference between the unweighted method and the adaptive method are usually within reasonable error margin. Overall, the key observation is that our proposed method provides better prediction than the KQR except the ozone data when $\tau = 0.8$. When analyzed using the complete ozone data, the only exception arises from the fact that all predictors are chosen at 80% quantile. Thus, it will not be too unexpected our method will be less efficient than the KQR in a non-sparse structure.

Table 4: Estimated prediction risk for two real datasets. The standard error is given in the parentheses.

τ	Method	Prostate	Ozone	WIPP
0.20	Unweighted SNQR	0.241 (0.007)	1.118 (0.016)	3278.731 (54.314)
	Adaptive SNQR	0.231 (0.007)	1.119 (0.016)	3183.783 (54.896)
	KQR-5CV	0.250 (0.009)	1.149 (0.017)	3378.978 (53.197)
0.50	Unweighted SNQR	0.318 (0.008)	1.660 (0.024)	5446.351 (85.819)
	Adaptive SNQR	0.334 (0.008)	1.685 (0.024)	5383.629 (83.122)
	KQR-5CV	0.342 (0.009)	1.707 (0.022)	5437.336 (83.375)
0.80	Unweighted SNQR	0.219 (0.006)	1.181 (0.017)	4639.145 (90.138)
	Adaptive SNQR	0.222 (0.006)	1.184 (0.017)	4413.725 (78.351)
	KQR-5CV	0.324 (0.011)	1.156 (0.016)	5047.170 (90.612)

Apart from comparing prediction error, we also apply our methods to the complete prostate data and summarize variable selection. An interesting comparison is that in the study of mean function, Tibshirani (1996) selected three prominent predictors, log-cancer volume, log-weight and seminal vesicle invasion. These three predictors are also selected by our approach when we consider the median. However, in the 20% quantile, gleason score shows up as an additional predictor. Meanwhile, in the 80% quantile, only two predictors are chosen, log-cancer volume and seminal vesicle invasion, but not log-weight.

6. CONCLUSIONS

We propose a new regularization method that simultaneously select important predictors and estimate conditional quantile function. Our SNQR method conquers the limitation of selecting only predictors that influence the conditional mean in least square regression, facilitating the analysis heteroskedastic data. The proposed method also includes the L_1-norm quantile regression and the KQR as special cases. In a simulation study and real data analysis, our method provides satisfactory model fitting and great potential for selecting important predictors.

The number of predictors we consider in both simulation and real data is moderate. With advancement of modern technology, high-throughput data becomes more frequent nowadays. In ultra-high dimensional feature space, Fan, Feng and Song (2011) recently proposed a screening procedure for nonparametric regression model. Further study can work toward incorporating a suitable screening procedure as a first step and then apply our proposed method at the second in a ultra-high dimensional feature space.

7. APPENDIX

Proof of Solution Existence

Proof. Denote the function to be minimized in (11) by

$$A(f) = n^{-1}\sum_{i=1}^{n}\rho_\tau(y_i - f(\boldsymbol{x}_i)) + \lambda J(f)$$

Without loss of generality, let $w_j = 1, \forall j$ and $\lambda = 1$. By decomposition in (9), for any $f \in \mathcal{F}_1$, we have $||f|| = ||\sum_{j=1}^{d} P^j f|| \leq \sum_{j=1}^{d}||P^j f|| = J(f)$. Denote the reproducing kernel and inner product of $\mathcal{F}_1$ be $R_{\mathcal{F}_1}(\cdot,\cdot)$ and $\langle\cdot,\cdot\rangle_{\mathcal{F}_1}$. By the definition of reproducing kernel,

$$\begin{aligned}|f(\boldsymbol{x}_i)| &= |\langle f(\cdot), R_{\mathcal{F}_1}(\boldsymbol{x}_i,\cdot)\rangle_{\mathcal{F}_1}| \leq \sqrt{\langle f(\cdot), f(\cdot)\rangle}\sqrt{\langle R_{\mathcal{F}_1}(\boldsymbol{x}_i,\cdot), R_{\mathcal{F}_1}(\boldsymbol{x}_i,\cdot)\rangle_{\mathcal{F}_1}} \\ &= ||f||\sqrt{R_{\mathcal{F}_1}(\boldsymbol{x}_i,\boldsymbol{x}_i)} \leq a||f|| \leq aJ(f)\end{aligned}$$

where $a^2 = \max_{i=1}^{n} R_{\mathcal{F}_1}(\boldsymbol{x}_i,\boldsymbol{x}_i)$ and the first inequality holds by Cauchy-Schwarz inequality. Denote $\rho = \max_{i=1}^{n}|y_i|$. Consider the set

$$D(f) = \left\{f \in \mathcal{F} : f = b + f_1, b \in \{1\}, f_1 \in \mathcal{F}_1, J(f) \leq \rho, |b| \leq (\min\{\tau,(1-\tau)\}^{-1} + a + 1)\rho\right\}$$

Then D is a closed, convex and bounded set. By Theorem 4 of Tapia and Thompson (1978), there exists a minimizer of (11) in D. Denote the minimizer by $\bar{f}$. Since a constant function $f(\boldsymbol{x}) = y_{([n\tau])}$, the sample $100\tau\%$ quantile, is also in D, we have $A(\bar{f}) \leq A(y_{[n\tau]}) \leq \rho$. Conversely, if $f \notin D$, then it is either (i) $J(f) > \rho$, or (ii) $|b| > (\min\{\tau,(1-\tau)\}^{-1} + a + 1)\rho$. In case (i), we have $A(f) \geq J(f) > \rho$. Whereas in the second case, we first notice

$$\begin{aligned}\rho_\tau(y_i - b - f_1) &\geq \min\{\tau, 1-\tau\}|b - (y_i - f_1)| \geq \min\{\tau, 1-\tau\}\{|b| - |y_i| - |f_1|\} \\ &> \min\{\tau,(1-\tau)\}\{(\min\{\tau,(1-\tau)\}^{-1} + a + 1)\rho - \rho - a\rho\} = \rho\end{aligned}$$

thus $A(f) > \rho$. Thus, for either case, we have $A(f) > A(\bar{f})$, that is $\bar{f}$ is a minimizer of (11).

□

Proof of Representer Theorem

Proof. Without loss of generality, let $w_j = 1$, $\forall j$. For any $f \in \mathcal{F}$, by decomposition in (6), we can write $f = b + \sum_{i=1}^{d} f_j$, where $f_j \in \mathcal{F}_j$. Denote g_j as the projection of f_j onto the space spanned by $R_j(\cdot,\cdot)$ and h_j as its orthogonal complement. Then $f_j = g_j + h_j$ and $||f_j||^2 = ||g_j||^2 + ||h_j||^2$. Since the reproducing kernel of $\mathcal{F}$ is $1 + \sum_{j=1}^d R(\cdot,\cdot)$, by reproducing theorem, we have

$$\begin{aligned} f(x_i) &= \left\langle 1 + \sum_{j=1}^d R(x_i,), b + \sum_{j=1}^d f_j \right\rangle \\ &= \left\langle 1 + \sum_{j=1}^d R(x_i,), b + \sum_{j=1}^d (g_j + h_j) \right\rangle \\ &= b + \left\langle 1 + \sum_{j=1}^d R(x_i,), b + \sum_{j=1}^d g_j \right\rangle + \left\langle 1 + \sum_{j=1}^d R(x_i,), b + \sum_{j=1}^d h_j \right\rangle \\ &= b + \sum_{j=1}^d \langle R(x_i,), g_j \rangle \end{aligned}$$

where $\langle \cdot,\cdot \rangle$ is the inner product in $\mathcal{F}$.

By substituting above expression into (11), the objective function becomes

$$n^{-1} \sum_{i=1}^n \rho_\tau \left(y_i - b - \sum_{j=1}^d \langle R_j(x_i,), g_j \rangle \right) + \lambda \sum_{j=1}^d \left(||g_j||^2 + ||h_j||^2 \right)^{1/2}$$

As a result, the minimizer should be chosen such that $||h_j||^2 = 0$ and therefore completes the proof. □

Derivation of Quadratic Programming Formula

Proof. To solve (16), we firs introduce slack variables $r_i^+ = (y_i - b - \sum_{k=1}^n c_k R_{ki}^\theta)_+$ and $r_i^- = (y_i - b - \sum_{k=1}^n c_k R_{ki}^\theta)_-$, then write the optimization problem in (16) in a matrix form

$$\begin{aligned} \min \quad & \tau \mathbf{1}^T \boldsymbol{r}^+ + (1-\tau)\mathbf{1}^T \boldsymbol{r}^- + n\lambda_0 \boldsymbol{c}^T \boldsymbol{R}^\theta \boldsymbol{c} \\ \text{s.t.} \quad & \boldsymbol{r}^+ \geq 0,\ \boldsymbol{r}^- \geq 0,\ b\mathbf{1} + \boldsymbol{R}^\theta \boldsymbol{c} + \boldsymbol{r}^+ - \boldsymbol{r}^- - \boldsymbol{y} = \mathbf{0} \end{aligned}$$

Then the foregoing setting gives the Lagrange primal function,

$$\mathcal{L} = \tau \mathbf{1}^T \boldsymbol{r}^+ + (1-\tau)\mathbf{1}^T \boldsymbol{r}^- + n\lambda_0 \boldsymbol{c}^T \boldsymbol{R}^\theta \boldsymbol{c} + \boldsymbol{\lambda}_1^T [b\mathbf{1} + \boldsymbol{R}^\theta \boldsymbol{c} + \boldsymbol{r}^+ - \boldsymbol{r}^- - \boldsymbol{y}] - \boldsymbol{\lambda}_2^T \boldsymbol{r}^+ - \boldsymbol{\lambda}_3^T \boldsymbol{r}^-$$

where $\boldsymbol{\lambda}_1 \in \mathbb{R}^n, \boldsymbol{\lambda}_2 \geq \mathbf{0}, \boldsymbol{\lambda}_3 \geq \mathbf{0}$ are Lagrange multipliers. By differentiating $\mathcal{L}$ with respect to $\boldsymbol{r}^{+}$, $\boldsymbol{r}^{-}$, b and $\boldsymbol{\theta}$, we arrive at

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{r}^{+}} &: \ \tau \mathbf{1} + \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2 = \mathbf{0} \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{r}^{-}} &: \ (1-\tau)\mathbf{1} - \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_3 = \mathbf{0} \\
\frac{\partial \mathcal{L}}{\partial b} &: \ \boldsymbol{\lambda}_1^T \mathbf{1} = 0 \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{c}} &: \ 2n\lambda_0 \boldsymbol{R}^{\theta}\boldsymbol{c} + \boldsymbol{R}^{\theta}\boldsymbol{\lambda}_1 = \mathbf{0}
\end{aligned}$$

By substituting these conditions into the Lagrange primal function, the dual problem is given by

$$\min_{\boldsymbol{c}} -\boldsymbol{y}^T\boldsymbol{c} + \frac{1}{2}\boldsymbol{c}^T\boldsymbol{R}^{\theta}\boldsymbol{c}$$

subject to the constraints

$$\mathbf{1}^T\boldsymbol{c} = 0, \ \frac{\tau-1}{2n\lambda_0}\mathbf{1} \leq \boldsymbol{c} \leq \frac{\tau}{2n\lambda_0}\mathbf{1}$$

□

Derivation of Linear Programming Formula

Proof. To solve (17), we take similar route as solving (16) by introducing slack variables $e_i = |y_i^* - \sum_{j=1}^{d}\theta_j G_{ij}|, e_i^+ = (y_i^* - \sum_{j=1}^{d}\theta_j G_{ij})_+$ and $e_i^- = (y_i^* - \sum_{j=1}^{d}\theta_j G_{ij})_-$ and re-write the objective function into a matrix form

$$\begin{aligned}
\tau\mathbf{1}^T\boldsymbol{e}^+ + (1-\tau)\mathbf{1}^T\boldsymbol{e}^- + n\lambda_0\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{\theta} &= \mathbf{1}^T\boldsymbol{e}^- + \tau\mathbf{1}^T(\boldsymbol{e}^+ - \boldsymbol{e}^-) + n\lambda\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{\theta} \\
&= \frac{1}{2}\mathbf{1}^T\boldsymbol{e} - \frac{1}{2}\mathbf{1}^T(\boldsymbol{e}^+ - \boldsymbol{e}^-) + \tau\mathbf{1}^T(\boldsymbol{y}^* - \boldsymbol{G}\boldsymbol{\theta}) + n\lambda\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{\theta} \\
&= \frac{1}{2}\mathbf{1}^T\boldsymbol{e} + \left(\tau - \frac{1}{2}\right)\mathbf{1}^T(\boldsymbol{y}^* - \boldsymbol{G}\boldsymbol{\theta}) + n\lambda_0\boldsymbol{c}^T\boldsymbol{G}\boldsymbol{\theta}
\end{aligned}$$

Since $\left(\tau - \frac{1}{2}\right)\mathbf{1}^T\boldsymbol{y}^*$ is a constant, the objective function can be simplified as

$$\min_{\boldsymbol{\theta},\boldsymbol{e}} \begin{pmatrix} n\lambda_0\boldsymbol{c}^T\boldsymbol{G} - (\tau - 0.5)\mathbf{1}^T\boldsymbol{G} & \frac{1}{2}\mathbf{1}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{e} \end{pmatrix}$$

subject to the constraints

$$\mathbf{1}^T\boldsymbol{\theta} \leq M, \ \theta_j \geq 0, \ \forall j, \quad -\boldsymbol{G}\boldsymbol{\theta} - \boldsymbol{e} \leq \boldsymbol{y}^*, \ \ \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{e} \geq \boldsymbol{y}^*\boldsymbol{e}$$

□

REFERENCES

Breiman, L. (1995), "Better subset selection using the nonnegative garrote," *Technometrics*, 37, 373–384.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear smoothers and additive models (with discussion)," *Annals of Statistics*, 17, 453–555.

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric independence screening in sparse ultra-high dimensional additive model," *Journal of the American Statistical Association*, 106, 544–557.

Gu, C. (2002), *Smoothing Spline ANOVA Models* New York: Springler-Verlag.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models* London: Chapman and Hall.

He, X., and Ng, P. (1999), "Quantile splines with several covariates," *Journal of Statistical Planning and Inference*, 75, 343–352.

He, X., Ng, P., and Portnoy, S. (1998), "Bivariate quantile smoothing splines," *Journal of the Royal Statistical Society, Ser. B*, 60, 537–550.

He, X., and Shi, P. (1994), "Convergence rate of B-spline estimators of nonparametric quantile functions," *Journal of Nonparametric Statistics*, 3, 299–308.

Jurečková, J., and Procházka, B. (1994), "Regression quantiles and trimmed least squares estimator in nonlinear regression model," *Journal of Nonparametric Statistics*, 3, 201–222.

Koenker, R. (2005), *Quantile Regression* New York: Cambridge University Press.

Koenker, R., and Bassett, G. (1978), "Regression quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and Hallock, K. (2001), "Quantile Regression," *Journal of Economic Perspectives*, 15, 143–156.

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile smoothing splines," *Biometrika*, 81, 673–680.

Li, Y., Liu, Y., and Zhu, J. (2007), "Quantile regression in reproducing kernel Hilbert spaces," *Journal of the American Statistical Association*, 477, 255–267.

Li, Y., and Zhu, J. (2008), "L_1-norm quantile regression," *Journal of Computational and Graphical Statistics*, 17, 163–185.

Lin, Y., and Zhang, H. H. (2006), "Component selection and smoothing in smoothing spline analysis of variance model," *Annals of Statistics*, 34, 2272–2297.

Procházka, B. (1988), "Regression quantiles and trimmed least squares estimator in the nonlinear regression model," *Computational Statistics and Data Analysis*, 6, 385–391.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 1135–1151.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989), "Prostate specific antigen in the disgnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients," *Journal of Urology*, 16, 1076–1083.

Storlie, C., Bondell, H., Reich, B., and Zhang, H. H. (2011), "The adaptive COSSO for nonparametric surface estimation and model selection," *Statistica Sinica*, 21, 679–705.

Storlie, C., and Helton, J. (2008), "Multiple predictor smoothing methods for sensitivity analysis: Example results," *Reliability Engineering and System Safety*, 93, 55–77.

Tapia, R., and Thompson, J. (1978), *Nonparametric Probability Density Estimation* Baltimore: John Hopkins Univ. Press.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

Wahba, G. (1990), *Spline Models for Observational Data* Philadelphia: SIAM.

Wu, Y., and Liu, Y. (2009), "Variable selection in quantile regression," *Statistics Sinica*, 19, 801–817.

Zou, H. (2006), "The adaptive LASSO and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., and Yuan, M. (2008*a*), "Composite quantile regression and the oracle model selection theory," *Annals of Statistics*, 36, 1108–1126.

Zou, H., and Yuan, M. (2008*b*), "Regularized simultaneous model selection in multiple quantiles regression," *Computational Statistics and Data Analysis*, 52, 5296–5304.