Quantile regression feature selection and estimation with grouped variables using Huber approximation

Ben Sherwood[1*] and Shaobo Li[1]

[1*]School of Business, University of Kansas, 1654 Naismith Dr., Lawrence, 66045, KS, USA.

*Corresponding author(s). E-mail(s): ben.sherwood@ku.edu;
Contributing authors: shaobo.li@ku.edu;

Abstract

This paper considers model selection and estimation for quantile regression with a known group structure in the predictors. For the median case the model is estimated by minimizing a penalized objective function with Huber loss and the group lasso penalty. While, for other quantiles an M-quantile approach, an asymmetric version of Huber loss, is used which approximates the standard quantile loss function. This approximation allows for efficient implementation of algorithms which rely on a differentiable loss function. Rates of convergence are provided which demonstrate the potential advantages of using the group penalty and that bias from the Huber-type approximation vanishes asymptotically. An efficient algorithm is discussed, which provides fast and accurate estimation for quantile regression models. Simulation and empirical results are provided to demonstrate the effectiveness of the proposed algorithm and support the theoretical results.

Keywords: quantile regression; group lasso; Huber regression

1 Introduction

We consider the problem of model selection and estimation of a conditional quantile when there is a group structure to the predictors, such as non-binary categorical variables or polynomial transformations of a covariate. Our focus is

on computationally efficient estimation of this model that comes with strong theoretical guarantees. Quantile regression directly models the conditional quantiles by minimizing the quantile loss, which is the absolute value of errors for the median and a tilted absolute value, referred to as the check function, for other quantiles (Koenker and Bassett, 1978). The objective functions are non-differentiable, but the minimization process can be formulated as an efficient linear programming problem (Koenker and D'Orey, 1987; Portnoy and Koenker, 1997). However for large data sets estimates can be unstable and computationally costly. One solution, expectile regression, is to replace the quantile loss with a quadratic loss equivalent, but this does not estimate conditional quantiles and is not robust to heavy tailed errors (Newey and Powell, 1987).

Breckling and Chambers (1988) proposed M-quantiles which generalizes quantile and expectile regression with an asymmetric Huber-loss. Others have recognized the advantages of smoothing the quantile loss and we provide some examples from what is inevitably an incomplete list. Motivated by balancing the bias and variance in a quantile regression estimator, Lee et al (2012) proposed approximating the the quantile loss with squared error loss near the non-differentiable point. Muggeo et al (2012) proposed an iterative least squares approximation to derive an exact path-following algorithm. Fasiolo et al (2021) proposed a new smooth generalization of the quantile loss function along with statistical advantages of this loss function when estimating general additive quantile models. Yi and Huang (2017) proposed approximating the quantile loss function with the Huber loss to solve elastic net quantile regression problems to increase computational efficiency. While influenced by these works, our approach differs in several ways. First, in this paper we consider a group lasso penalty to select variables with known grouped structure. Second, we study statistical consistency for the penalized estimator and provide rates of convergence with diverging number of predictors, which was not studied in Yi and Huang (2017). Third, we provide a fast computing algorithm similar to Yang and Zou (2015), which is based on the majorize-minimize (MM) algorithm. Fourth, while we use an M-quantile loss function, the theoretical results are with respect to a population quantiles not a population M-quantile. In summary, our work proposes an efficient algorithm for simultaneous model selection and estimation of quantile regression with grouped predictors, while also providing statistical guarantees.

Since Breckling and Chambers (1988), there have been several papers that explored using M-quantiles as an alternative to quantile regression. The population version of an M-quantile is not equivalent to a population quantile but an often cited reason for preferring M-quantile estimators is computational convenience and stability (Chambers and Tzavidis, 2006; Tzavidis et al, 2016; Alfo et al, 2017). One popular use of M-quantiles is to estimate regression coefficients while accounting for the heterogeneity that often comes from clustered data. Examples include modeling air quality with mixture models (Del Sarto et al, 2019), small area estimation (Chambers and Tzavidis, 2006; Bianchi

et al, 2018) and modeling the emotional and behavioral disorders of UK children using a multivariate response longitudinal model (Tzavidis et al, 2016). The work presented here can be considered an M-quantile estimator with a group lasso penalty, which would be a new contribution to the M-quantile literature. The major difference between our work and the M-quantile literature is we prove consistency to quantile regression coefficients, not the biased M-quantiles.

Methods that provide simultaneous estimation and model selection have been an active area of interest since the seminal paper by Tibshirani (1996). One important extension of this work is the group lasso which performs simultaneous estimation and model selection, while guaranteeing that groups of variables, such as polynomial functions of a continuous variable or dummy variables from a categorical variable, to be selected as a group (Yuan and Lin, 2005). The statistical properties of group lasso estimators have been studied in a variety of setting including linear mean regression (Huang and Zhang, 2010; Lounici et al, 2011; Negahban et al, 2012), logistic regression (Meier et al, 2008), nonparametric additive models (Huang et al, 2010) and quantile regression (Kato, 2011; Ciuperca, 2019). This work differs from the quantile regression approaches because we approximate the quantile loss function with the Huber loss function, which allows for faster computation, while asymptotic results remain the same. Huber loss with the group lasso penalty has been used for consistent model selection across multiple data sets to robustly estimate gene regulatory networks (Liu et al, 2014). Our work is different because we consider the Huber loss function for quantile regression and provide rates of convergence.

Robust regression using the Huber loss was first considered by Huber (1973). Asymptotic results were derived assuming the parameter of interest minimizes the expected Huber loss, which will be satisfied when the distribution of the errors are symmetric. In addition, the tuning parameter controlling the balance between squared and absolute error loss was considered fixed. These assumptions have been relaxed and non-asymptotic results have been derived that allow for tuning parameter to change with the sample size and the errors to come from an asymmetric distribution (Zhou et al, 2018; Sun et al, 2020). The lasso penalty has been added to consider simultaneous model selection and estimation for robust models that can handle asymmetric errors (Fan et al, 2017; Zhou et al, 2018). Fan et al (2017), Zhou et al (2018) and Sun et al (2020) demonstrate that Huber loss can provide a consistent estimator of mean regression coefficients, when the loss function becomes a closer approximation to the squared error loss when more data is collected. This paper demonstrates that asymmetric Huber loss with a group penalty can provide consistent estimators of quantile regression coefficients when the approximation becomes closer to the quantile loss when more data is collected. While their work emphasized the robust advantages of Huber loss over squared error loss, our work emphasizes the computational benefits of asymmetric Huber loss over quantile loss.

Quantile loss and the lasso penalty functions have similar forms, identical in the case of median regression, and minimizing quantile loss plus the lasso penalty can also be framed as a linear programming problem (Xu and Ying, 2010; Wu and Liu, 2009; Belloni and Chernozhukov, 2011). Many penalized quantile regression problems can be solved using convex optimization software (Koenker and Mizera, 2014). However, for high-dimensional problems the linear programming solutions can be computationally inefficient. When the lasso penalty is replaced with a group lasso penalty the problem can be framed as a second order cone programming problem, but these tend to be even slower than linear programming problems. Yi and Huang (2017) demonstrated that replacing the check function with Huber loss can greatly increase computational efficiency, while sacrificing little with respect to estimation accuracy, for fitting elastic-net penalized quantile regression. Huber loss is differentiable, unlike the check function, which allows for the application of penalized regression algorithms that depend on differentiable loss functions. Yi and Huang (2017) used a semismooth coordinate descent algorithm, but this approach does not directly apply to the group lasso. Instead, we apply the algorithm proposed by Yang and Zou (2015) for group lasso problems with a differentiable loss function.

In Section 2 we introduce the method and provide rates of convergence for the estimators. Details about the algorithm are provided in Section 3. Simulations will be provided in Section 4 and an application to the Ames housing data set (De Cock, 2011) will be provided in Section 5.

1.1 Notation

This subsection introduces notations that will be used throughout the paper. Vectors, such as $\mathbf{a} \in \mathbb{R}^p$, will be represented by bold lowercase letters, while matrices, such as $A \in \mathbb{R}^{n\times p}$, will be represented by italics capital letters. For any vector $\|\mathbf{a}\|_q$ represents the L^q-norm and $\|\mathbf{a}\|_\infty$ is the maximum element of the vector $\mathbf{a}$. For any matrix A, $\|A\|$ is the spectral norm. For subspace $\mathcal{A} \subseteq \mathbb{R}^p$, the vector $\mathbf{a}_\mathcal{A}$ is the projection of $\mathbf{a}$ to $\mathcal{A}$ and $\mathcal{A}^\perp$ is the orthogonal complement of $\mathcal{A}$. A random variable is $o_P(1)$ if it converges in probability to zero and $o(1)$ represents that a deterministic sequence converges to zero. For random variable X_n and distribution F, $X_n \xrightarrow{d} F$ represents that X_n converges in distribution to F. For a scalar $a \in \mathbb{R}$, $(a)_+ = \max(0, a)$.

2 Method

Consider independent identically distributed random variables $\{y_i, \mathbf{x}_i\}_{i=1}^n$ with $y_i \in \mathbb{R}$ and $\mathbf{x}_i \in \mathbb{R}^{p+1}$. The predictors, $\mathbf{x}_i$, are partitioned into G groups. Define $\mathbf{x}_{ig} \in \mathbb{R}^{d_g}$ as the vector of the d_g elements of $\mathbf{x}_i$ that belong to the gth group. Without loss of generality assume that the data is organized by the groups of predictors, that is, $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^\top = (1, \mathbf{x}_{i1}^\top, \ldots, \mathbf{x}_{iG}^\top)^\top$, where $x_{i0} = 1$, but we use x_{i0} throughout the paper for notational convenience.

Let $\tau \in (0,1)$ be the quantile of interest. The linear quantile regression model is

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_\tau^* + \epsilon_i^\tau, \tag{1}$$

where $\boldsymbol{\beta}_\tau^* = (\beta_{\tau,0}^*, \beta_{\tau,1}^*, \ldots, \beta_{\tau,p}^*)^\top = (\beta_{\tau,0}^*, \boldsymbol{\beta}_{\tau,1}^{*\top}, \ldots, \boldsymbol{\beta}_{\tau,G}^{*\top})^\top \in \mathbb{R}^{p+1}$ and $P(\epsilon_i^\tau < 0 \mid \mathbf{x}_i) = \tau$. Define $\rho_\tau(u) = u[\tau - I(u < 0)]$, which is the quantile loss for the τth quantile. Note, $\boldsymbol{\beta}_\tau^* = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} E[\rho_\tau(y - \mathbf{x}^\top \boldsymbol{\beta})]$ and therefore is typically estimated by minimizing $\sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ (Koenker and Bassett, 1978).

The quantile loss function has the equivalent definition of $\rho_\tau(u) = \frac{1}{2}[|u| + (2\tau - 1)u]$, where $|u|$ can be approximated by the Huber loss function (Huber, 1964), which is of form

$$h_\gamma(t) = \begin{cases} \dfrac{t^2}{2\gamma}, & \text{if } |t| \le \gamma, \\ |t| - \dfrac{\gamma}{2}, & \text{if } |t| > \gamma. \end{cases}$$

Therefore, we define the Huber-approximated quantile loss as

$$h_\gamma^\tau(u) = h_\gamma(u) + (2\tau - 1)u. \tag{2}$$

For sufficiently small γ then $\rho_\tau(u) \approx h_\gamma^\tau(u)/2$ and the solution that minimizes $\sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$ will be similar to the minimizer of $\frac{1}{2}\sum_{i=1}^n h_\gamma^\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$. Specifically, consider the objective function of $H_\gamma^\tau(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^n h_\gamma^\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})$ and estimator of

$$\hat{\boldsymbol{\beta}}_\gamma^\tau = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} H_\gamma^\tau(\boldsymbol{\beta}). \tag{3}$$

The smaller the value of γ the closer (3) is to the standard quantile regression objective function, and for the special case of $\gamma = 0$ they are identical. Figure 1 illustrates the approximation of the quantile loss $\rho_\tau(u)$ using the Huber-approximated quantile loss, $h_\gamma^\tau(u)/2$, for $\tau \in \{0.25, 0.5, 0.75\}$ and $\gamma \in \{0.01, 0.1, 0.2\}$.

The estimator $\hat{\boldsymbol{\beta}}_\gamma^\tau$ was first proposed by Breckling and Chambers (1988) as an M-quantile estimator to provide a generalized asymmetric error loss that had expectile regression and quantile regression as special cases. Kokic et al (1997) demonstrated that M-quantiles are guaranteed to have a unique minimizer, a property that quantile regression does not have. Previous theoretical results prove asymptotic results with respect to the true conditional M-quantile (Breckling and Chambers, 1988; Bianchi and Salvati, 2015; Pratesi et al, 2009). The results in this paper differ because asymptotic results are established with respect to the true quantile regression coefficients.

The next two subsections will consider the fixed and high dimensional setting. Results in Section 2.1 will demonstrate that using $\hat{\boldsymbol{\beta}}_\gamma^\tau$ is theoretically

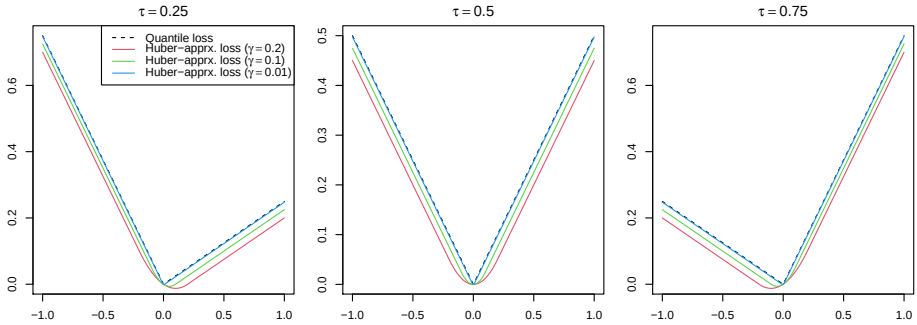

Fig. 1 Quantile loss (dashed line) vs. Huber-approximated quantile loss (solid line).

justified because for $\gamma = o(1)$ it is asymptotically equivalent to the standard quantile regression estimator. In Section 2.2, a group penalty will be added to (3) to perform simultaneous estimation and model selection, while incorporating the group structure of the predictors.

2.1 Fixed dimension case

The following conditions were used to evaluate the asymptotic properties of $\hat{\boldsymbol{\beta}}_{\gamma}^{\tau}$.

Condition 1 The conditional cumulative distribution functions, $F_i(\cdot \mid \mathbf{x}_i)$, of the error terms are absolutely continuous. For the conditional probability density function, $f_i(\cdot \mid \mathbf{x}_i)$, there exists positive constants c_f, C_f and C'_f such that for all $i \in \{1, \ldots, n\}$ $c_f \leq f_i(0 \mid \mathbf{x}_i) \leq C_f$ and in a neighborhood around zero $|f'(\cdot \mid \mathbf{x}_i)| \leq C'_f$.

Condition 2 There exists a positive constant C_x such that $|x_{ij}| < C_x$ for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, p\}$. Also, there exists positive definite matrices Σ_1 and Σ_2 such that

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\top}\right) = \Sigma_1 \text{ and } E\left(\frac{1}{n}\sum_{i=1}^{n}f_i(0 \mid \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^{\top}\right) = \Sigma_2.$$

Conditions 1 and 2 are common conditions for quantile regression and are similar to conditions used in Section 4.2 of Koenker (2005) for the proof of asymptotic normality of the standard quantile regression estimator. Our results do not depend on a moment condition for the error terms unlike recent Huber results for estimating mean regression coefficients (Fan et al, 2017; Sun et al, 2020; Zhou et al, 2018) and M-quantile theory (Breckling and Chambers, 1988; Bianchi and Salvati, 2015; Pratesi et al, 2009). The following theorem demonstrates that for $\gamma \to 0$ the estimator $\hat{\boldsymbol{\beta}}_{\gamma}^{\tau}$ is asymptotically equivalent to the standard quantile regression estimator.

Theorem 1 *Assume Conditions 1-2 hold, p is a fixed constant and $\gamma = o(1)$ then*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\gamma}^{\tau} - \boldsymbol{\beta}_{\tau}^{*}\right) \xrightarrow{d} N[\mathbf{0}, \tau(1-\tau)\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}].$$

Proof Proof of Theorem 1 and all other theoretical results are available in the supplemental material. □

Theorem 1 demonstrates that while replacing the quantile loss with a Huber approximation introduces some bias in the estimation, for the correct sequence of γ the bias is negligible asymptotically. There exist asymptotic results in the M-quantile literature for fixed γ but in that setting the estimator converges to the true M-quantile coefficients, which are not necessarily the same as the true quantile regression coefficients (Breckling and Chambers, 1988; Bianchi and Salvati, 2015; Pratesi et al, 2009). The next section explores if these results hold for high-dimensional estimators.

2.2 High dimensional Case

Assume, without loss of generality, that only the coefficients of the first q groups have non-zero values. That is, for $d^* = \sum_{g=1}^{q} d_g$ we have $\boldsymbol{\beta}_{\tau}^{*} = \left(\beta_0, \boldsymbol{\beta}_1^{\top}, \ldots, \boldsymbol{\beta}_q^{\top}, \mathbf{0}_{p+1-d^*}^{\top}\right)^{\top}$. To simultaneously perform estimation and variable selection while accounting for the group structure in the predictors we consider the penalized objective function of $H_{\gamma,\lambda}^{\tau}(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n} h_{\gamma}^{\tau}(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}) + \lambda\sum_{g=1}^{G}\sqrt{d_g}\|\boldsymbol{\beta}_g\|_2$ with corresponding estimator

$$\hat{\boldsymbol{\beta}}_{\gamma,\lambda}^{\tau} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} H_{\gamma,\lambda}^{\tau}(\boldsymbol{\beta}). \tag{4}$$

Yi and Huang (2017) had a similar estimator for the conditional quantile, but used the elastic net penalty instead of the group lasso penalty. They focused on the computational advantages of using the Huber loss function over the quantile loss function and did not consider rates of convergence. Kato (2011) proposed an estimator which directly uses the quantile loss instead of the Huber approximation. That estimator can be stated as a second-order cone programming and can be solved using convex optimization software, see Koenker and Mizera (2014). In the following we demonstrate that the proposed estimator has similar asymptotic results to the results of Kato (2011). Then in the next two sections we demonstrate the computational speed gains of using the proposed estimator, while maintaining a similar performance to an estimator that uses the quantile loss function.

Unlike Theorem 1, the following conditions and results are for high dimensional estimators including the case where $G > n$. Under the assumption of sparsity the parameter space for $\boldsymbol{\beta}_{\tau}^{*}$ is $\mathcal{M} = \{\boldsymbol{\beta} \in \mathbb{R}^{p+1} \mid \boldsymbol{\beta}_g = \mathbf{0}_{d_g} \text{for all } g \in$

$\{q+1,\ldots,G\}\}$. Define $\|\mathbf{u}\|_{G,2} = \sum_{g=1}^{G} \sqrt{d_g}\|\mathbf{u}_g\|_2$ and

$$\mathcal{C}^{\tau}_{\gamma,\lambda} = \{\mathbf{u} \in \mathbb{R}^{p+1} \mid \|\mathbf{u}_{\mathcal{M}^{\perp}}\|_{G,2} \leq 3\|\mathbf{u}_{\mathcal{M}}\|_{G,2} + \gamma/\lambda\}. \tag{5}$$

Define

$$\Lambda_{\tau} = \max_{0\leq g\leq G} \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{ig} d_g^{-1/2} \left[\tau - I(\epsilon_i^{\tau} \leq 0)\right] \right\|_2, \tag{6}$$

which is the dual of group penalty of the subgradient of the quantile loss function evaluated at $\boldsymbol{\beta}^*_{\tau}$. Generally this dual is of interest for high-dimensional M-estimators because when it is upper bounded by some factor of λ it guarantees that the error of the estimator is in a cone similar to $\mathcal{C}^{\tau}_{\gamma,\lambda}$. The following lemma demonstrates this property holds for the proposed estimator.

Lemma 1 If $\lambda \geq 2\Lambda_{\tau}$ then for any $\boldsymbol{\beta}$ such that $H^{\tau}_{\gamma,\lambda}(\boldsymbol{\beta}) \leq H^{\tau}_{\gamma,\lambda}(\boldsymbol{\beta}^*_{\tau})$ it follows that $\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\tau} \in \mathcal{C}^{\tau}_{\gamma,\lambda}$.

Lemma 1 is similar to established results for norm-based regularized M-estimators that demonstrate with the correct choice of λ, the error between the estimator and the truth lies in a cone similar to $\mathcal{C}^{\tau}_{\gamma,\lambda}$. There exist general results for differentiable loss functions (Negahban et al, 2012) and results specific to quantile regression (Belloni and Chernozhukov, 2011; Kato, 2011). The key difference between existing results and Lemma 1 is the addition of the γ/λ term. This bias term arrives because of the bias introduced by using the Huber loss and that Λ_{τ} is defined with respect to quantile loss, not Huber loss. However, with γ/λ converging to zero we can apply Lemma 1 to establish a bound for $\|\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda} - \boldsymbol{\beta}^*_{\tau}\|_2$. First, we provide some conditions that will be used to derive the bound.

Condition 3 There exists positive constants $\phi_{\min}$ and $\phi_{\max}$ such that

$$0 < \phi_{\min} = \inf_{\mathbf{a}\in\mathcal{C}^{\tau}_{\gamma,\lambda}, \|\mathbf{a}\|_2=1} \|\Sigma_1^{1/2}\mathbf{a}\|_2 \leq \sup_{\mathbf{a}\in\mathcal{C}^{\tau}_{\gamma,\lambda}, \|\mathbf{a}\|_2=1} \|\Sigma_1^{1/2}\mathbf{a}\|_2 = \phi_{\max} < \infty.$$

Condition 4 For all $g \in \{1,\ldots,G\}$, $E[\mathbf{x}_g\mathbf{x}_g^{\top}] = I_{d_g}$.

Condition 5 There exists a positive constant C_r such that

$$0 < C_r = \frac{c_f}{C'_f\phi_{\max}} \inf_{\boldsymbol{\alpha}\in\mathcal{C}^{\tau}_{\gamma,\lambda}, \|\boldsymbol{\alpha}\|_2=1} \frac{E[(\boldsymbol{\alpha}^{\top}\mathbf{x})^2]^{3/2}}{E[|\boldsymbol{\alpha}^{\top}\mathbf{x}|^3]}.$$

Conditions 3-5 are similar to those used in Kato (2011). Condition 3 is a restricted eigenvalue condition which are very common in high-dimensional analysis. Our version of this restricted eigenvalue is slightly different due to

the definition of $\mathcal{C}^{\tau}_{\gamma,\lambda}$. However, our asymptotic results require $\gamma/\lambda \to 0$ thus making our condition asymptotically equivalent to the widely used restricted eigenvalue condition first presented in Bickel et al (2009). Condition 4 provides a nice structure to the covariates within a group and can be achieved by scaling the data in each group. Condition 5 is used to derive a lower bound for $E[\rho_\tau(y - \mathbf{x}^\top\boldsymbol{\beta}) - \rho_\tau(y - \mathbf{x}^\top\boldsymbol{\beta}^*_\tau)]$ when $\boldsymbol{\beta} - \boldsymbol{\beta}^*_\tau \in \mathcal{C}^{\tau}_{\gamma,\lambda}$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*_\tau\|_2$ is set to some known value. A similar condition was first introduced in Belloni and Chernozhukov (2011) for quantile regression with the lasso penalty and extended to quantile regression group lasso in Kato (2011).

Define $d_{\min} = \min_{g\in\{1,\ldots,G\}} d_g$ and $d_{\max} = \max_{g\in\{1,\ldots,G\}} d_g$. Define the event $\Omega_0 = \{\|\hat{\Sigma}_g - I_{d_g}\| \leq .5, \text{ for all } g \in \{1,\ldots,G\}\}$ and let $\kappa \in (0,1)$ be the value such that $P(\Omega_0) = 1 - \kappa$.

Theorem 2 *Let C_1, C_2 and C_3 be positive values. Define*

$$\delta^* = \frac{C_1\sqrt{\log(G)}16\sqrt{10}(4\sqrt{d^*+1}+1)}{\sqrt{nd_{\min}}} + \frac{8\sqrt{3}(4\sqrt{d^*+1}+1)}{\sqrt{n}}.$$

Assume Conditions 1 - 5 hold, $\lambda = \frac{2(4\sqrt{2}+C_2)}{\sqrt{n}} + 2C_3\sqrt{\frac{\log(G)}{nd_{\min}}}$ and $\gamma < \frac{3\lambda}{c_f\phi^2_{\min}}\left[\max\left(\delta^*, \sqrt{\frac{8\phi^2_{\max}}{n}}\right) + \lambda(\sqrt{d^*}+1/2)\right]$. Then with probability at least $1 - 2exp(-C_2^2/2) - 16G^{1-C_3^2/128} - 64G^{1-C_1^2} - 4\kappa$

$$\|\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda} - \boldsymbol{\beta}^*_\tau\|_2 \leq \frac{3}{c_f\phi^2_{\min}}\left[\max\left(\delta^*, \sqrt{\frac{8\phi^2_{\max}}{n}}\right) + \lambda(\sqrt{d^*}+1/2)\right],$$

so long as for sufficiently large n the provided upper bound is smaller than C_r.

Kato (2011) provided a similar result for the case of quantile regression using the quantile loss function. Thus this result demonstrates that for γ converging to zero at the right rate the estimator derived using the Huber approximation will have a similar error bound to an estimator derived using the quantile loss. The last condition regarding C_r dominating the bound, implicitly bounds d^* and is found in similar quantile regression work (Kato, 2011; Belloni and Chernozhukov, 2011). If n increases sufficiently faster than $d_{\max}$ then by Condition 4 it is reasonable to assume that $\kappa \to 0$. Under this assumption, Theorem 2 can be used to derive rates of convergence for when G is fixed or G increases to infinity with n.

Corollary 1 Assume Conditions 1 - 5 hold, $\kappa \to 0$, $\lambda = \frac{2(4\sqrt{2}+t)}{\sqrt{n}} + 2t\sqrt{\frac{\log(G)}{nd_{\min}}}$ where $t \to \infty$ and $\gamma < \frac{3\lambda}{c_f\phi^2_{\min}}\left[\max\left(\delta^*, \sqrt{\frac{8\phi^2_{\max}}{n}}\right) + \lambda(\sqrt{d^*}+1/2)\right]$ then for G fixed

$$\left\|\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda} - \boldsymbol{\beta}^*_\tau\right\|_2 = O_P\left\{t\sqrt{\frac{d^*}{n}\left[1 + \frac{\log(G)}{d_{\min}}\right]}\right\}.$$

Corollary 2 Assume Conditions 1 - 5 hold, $\kappa \to 0$, $\log(G) \to \infty$, $\lambda = \frac{2(4\sqrt{2}+\sqrt{t})}{\sqrt{n}} + 2T\sqrt{\frac{\log(G)}{nd_{\min}}}$ where $t \to \infty$, $T > 8\sqrt{2}$ and $\gamma < \frac{3\lambda}{c_f \phi^2_{\min}} \left[\max\left(\delta^*, \sqrt{\frac{8\phi^2_{\max}}{n}}\right) + \lambda(\sqrt{d^*} + 1/2)\right]$ then

$$\left\|\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda} - \boldsymbol{\beta}^*_{\tau}\right\|_2 = O_P\left\{\sqrt{\frac{d^*}{n}\left[t + \frac{\log(G)}{d_{\min}}\right]}\right\}.$$

Corollaries 1 and 2 demonstrate that the rate of convergence for the group lasso estimator can be separated into the rate of convergence for estimating the non-sparse parameters and the rate provided by using the group lasso penalty. The rate $\sqrt{\frac{d^*}{n}}$ is the rate of convergence if only the active groups were used to fit the model (Wang et al, 2012; He and Shao, 2000). While, $\frac{d^* \log(G)}{d_{\min} n}$ reflects the uncertainty in selecting the active groups. Define $w = \|\boldsymbol{\beta}^*_{\tau}\|_0$, Belloni and Chernozhukov (2011) derived a rate of $\sqrt{\frac{w \log[\max(p,n)]}{n}}$ for quantile regression with lasso. Thus if the active groups consist mostly of non-zero coefficients then the group lasso penalty provides an improved rate of convergence, but if this is not the case then the group lasso estimator may perform worse than the lasso estimator. Similar bounds and conclusions have been found in other work that used a group lasso penalty (Huang and Zhang, 2010; Lounici et al, 2011; Negahban et al, 2012; Kato, 2011).

3 Algorithm

Yang and Zou (2015) proposed a general algorithm, called groupwise-majorization-descent (GMD), for solving group lasso penalized objective functions which under mild conditions converges to the right solution. Similar to the majorize-minimize (MM) algorithm (Hunter and Lange, 2004; Wu et al, 2010), a key step is finding a majorizing function that provides an upper bound to the original objective function. This majorizing function is also convex, and it is a strict upper bound except at one point. Then the solution to the original objective function can be found by iteratively minimizing the majorizing function. The GMD algorithm is applicable to loss functions that satisfy the quadratic majorization (QM) condition, which can be verified to hold for our Huber-approximated loss function $h^{\tau}_{\gamma}(u)$. Specifically the following need to hold, (1) $h^{\tau}_{\gamma}(u)$ is differentiable everywhere, and (2) let X be the design matrix and $H = \frac{2}{n\gamma} X^{\top} X$, then for any $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$

$$H^{\tau}_{\gamma}(\boldsymbol{\beta}) \leq H^{\tau}_{\gamma}(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\top} \nabla H^{\tau}_{\gamma}(\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^{\top} H (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}). \tag{7}$$

The first order derivative of our loss function $h^{\tau}_{\gamma}(u)$ is Lipschitz continuous with constant $1/\gamma$, i.e., $|h'^{\tau}_{\gamma}(u_1) - h'^{\tau}_{\gamma}(u_2)| \leq \frac{1}{\gamma}|u_1 - u_2|$. Therefore (7) holds by Lemma 1 in Yang and Zou (2015).

We first describe the GMD algorithm for a fixed value of λ and then for the solution path of a sequence of λ values. Define H_g as the submatrix of H corresponding to the gth group and define ξ_g as the largest eigenvalue of H_g. The following is our Huber approximation quantile (HAQ) algorithm to estimate $\hat{\beta}_{\gamma,\lambda}$ starting with an initial estimate of the zero vector.

1. Iterate through the following updates for $g \in \{0, \ldots, G\}$ until convergence is reached. In our implementation, the algorithm converges if $\|\tilde{\boldsymbol{\beta}}(\text{new}) - \tilde{\boldsymbol{\beta}}\|_\infty < 10^{-4}$.
   (a) For $g = 0$, update $\tilde{\beta}_0(\text{new}) = \left\{-\gamma \nabla \left[H_\gamma^\tau(\tilde{\boldsymbol{\beta}})\right]_0 + \beta_0\right\}$.
   (b) For $g \in \{1, \ldots, G\}$ update

$$\tilde{\boldsymbol{\beta}}_g(\text{new}) = \frac{1}{\xi_g}\left\{-\left[\nabla H_\gamma^\tau(\tilde{\boldsymbol{\beta}})\right]_g + \xi_g \tilde{\boldsymbol{\beta}}_g\right\}\left(1 - \frac{\lambda\sqrt{d_g}}{\|-\left[\nabla H_\gamma^\tau(\tilde{\boldsymbol{\beta}})\right]_g + \xi_g\tilde{\boldsymbol{\beta}}_g\|_2}\right)_+. \quad (8)$$

2. If the algorithm did not converge then set $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\text{new})$. Upon convergence, set $\hat{\boldsymbol{\beta}}_{\gamma,\lambda}^\tau = \tilde{\boldsymbol{\beta}}$.

Derivation of (8) can be found in the supplemental material. As discussed in Yang and Zou (2015), each iteration of the algorithm strictly decreases the penalized objective function, unless the algorithm has converged. Given that the penalized objective function is convex, the algorithm converges to the right answer.

To solve for a sequence of λ values we use the warm start method to derive initial estimates. Specifically, the estimation starts from the most sparse model, i.e., the largest λ, and the estimates at $\lambda^{(k)}$ are used as the initial value for the solution at $\lambda^{(k+1)}$, where $\lambda^{(k)} > \lambda^{(k+1)}$. To further improve the computational efficiency, the strong rule (Tibshirani et al, 2012) that prescreens the coefficients at each λ is adopted.

Let $\hat{\boldsymbol{\beta}}_{\gamma,\lambda^{(k)}}^\tau$ be the solution obtained for $\lambda^{(k)}$. The strong rule suggests that for each group g, the solution at $\lambda^{(k+1)}$ is likely to be zero if

$$\|\nabla_g H_\gamma^\tau[\hat{\boldsymbol{\beta}}_{\gamma,\lambda^{(k)}}^\tau]\|_2 < \sqrt{d_g}[2\lambda^{(k+1)} - \lambda^{(k)}]. \quad (9)$$

Then the solution at $\lambda^{(k+1)}$ can be computed based on the reduced covariates, $X_{\mathcal{A}}$, where $\mathcal{A}$ is the set of group indices that corresponds to the nonzero coefficients suggested by the strong rule. Application of the strong rule can incorrectly discard groups of predictors with nonzero coefficients. To protect against this, upon convergence for the new estimator, $\hat{\boldsymbol{\beta}}_{\gamma,\lambda^{(k+1)}}^\tau$, the following KKT condition is checked for each $g \in \mathcal{A}^C$,

$$\|\nabla_g H_\gamma^\tau[\hat{\boldsymbol{\beta}}_{\gamma,\lambda^{(k+1)}}^\tau]\|_2 \leq \lambda^{(k+1)}\sqrt{d_g}. \quad (10)$$

Below we outline the GMD algorithm for a sequence of λ, where the application of the strong rule (9) and checking of KKT conditions (10) are involved.

1. Initializing
   (a) Standardize covariates, and set an initial estimator $\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda^{(0)}}$ as zero except for the intercept, which is set to $Q_{\tau}(y)$, the τth quantile of the observed response variable.
   (b) Construct a sequence λ values, $(\lambda^{(0)}, \lambda^{(1)}, ..., \lambda^{(K)})$. In particular, set $\lambda^{(0)} = \lambda_{max} = \max_g \|\nabla_g H^{\tau}_{\gamma}(\hat{\boldsymbol{\beta}}^{(0)})\|_2/\sqrt{d_g}$, which by KKT conditions ensures all non-intercept coefficients are zero. Following Friedman et al (2010), set $\lambda_{min} = 0.001\lambda_{max}$ if $n > p$ and $\lambda_{min} = 0.01\lambda_{max}$ if $n \leq p$, and the sequence of λ values are generated with uniform spacing between λ_{max} and λ_{min} on the natural log scale.
2. For each $\lambda^{(k)}$, $k = 1, ..., K$, do the following.
   (a) Initialize $\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda^{(k)}} = \hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda^{(k-1)}}$. Find $\mathcal{A}$, the set of group indices that correspond to the nonzero coefficients suggested by the strong rule (9).
   (b) Apply the HAQ algorithm to the reduced dataset $\{Y, X_{\mathcal{A}}\}$ to solve for $\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda^{(k)}}(\mathcal{A})$ and setting all other coefficients to zero. If the set $\mathcal{A}$ is full, stop and return the solution $\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda^{(k)}}$, otherwise proceed to step 2(c).
   (c) For each $g \in \mathcal{A}^c$, check the KKT condition (10). If (10) holds for all $g \in \mathcal{A}^c$, stop and return to the solution $\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda^{(k)}}$. Otherwise, update the set $\mathcal{A}$ by including any group index g for which (10) does not hold, and repeat step 2(b).

3.1 Choice of tuning parameter γ

A good choice for γ will properly balance between the efficiency of the algorithm and the bias of the estimator. Larger values of γ will result in a more computational efficient estimator, but will also increase the bias. For smaller values of γ the estimator becomes less biased, but less computationally efficient. In the extreme case of $\gamma = 0$ the proposed algorithm is invalid because the loss function is no longer differentiable. For computationally stability there need to be a sufficient number of residuals that use the squared error loss in h^{τ}_{γ}. One could choose γ by cross-validation, but that increases the computational cost. Given one of the motivating factors for this work is reducing computational costs, we implement an ad hoc approach similar to Yi and Huang (2017) that requires little computational cost to adaptively choose γ across the solution path. Specifically, let $\mathbf{r}^{(k)} \in \mathbb{R}^n$ be the vector of the residuals at $\lambda^{(k)}$. We set the value of $\gamma^{(k+1)}$ for $\lambda^{(k+1)}$ as

$$\gamma^{(k+1)} = \min\{\gamma_{\max}, \max(\gamma_{\min}, Q_{0.1}(|\mathbf{r}^{(k)}|))\}.$$

The values of $\gamma_{\max}$ and $\gamma_{\min}$ specify the range of γ to be chosen, and $Q_{0.1}(|\mathbf{r}^{(k)}|)$ is the 10th percentile of absolute value of the residuals. The key idea is to get a value of γ that is not too large or small. Our implementation sets $\gamma_{\max} =$

4 and $\gamma_{\min} = 0.001$. One may supply an initial value γ_0 depending on the scale of the response variable. We recommend the initial value γ_0 to be in the range (IQR(y)/50, IQR(y)/100). In our implementation, we set $\gamma_0 = 0.2$ for simulation studies and $\gamma_0 = 0.005$ for the real data example. Based on our experiments, the choice of γ is relatively insensitive as long as it is within a proper range. In the next section we demonstrate that for our simulation settings performance of the proposed method is fairly robust to the choice of γ.

4 Simulations

In this section we conduct simulation studies to examine the effectiveness of the proposed method. The settings are similar to those presented in Yuan and Lin (2005), but modified for quantile instead of mean regression. Specifically, let $\tilde{X}_i = (Z_i + W)/\sqrt{2}$, where $(Z_1, ..., Z_p) \sim \text{MVN}(0, I_p)$ and $W \sim N(0, 1)$. Then let $X_i = \tilde{X}_i$ for $i = 1, ..., p_1$ and $X_i = I(\tilde{X}_i > z_*) + I(\tilde{X}_i > -z_*)$ for $i = p_1 + 1, ..., p$, where $z_* = \Phi^{-1}(2/3)$. That is, $X_1, ..., X_{p_1}$ are continuous and $X_{p_1+1}, ..., X_p$ are trichotomized as 0, 1 or 2. The response variable is generated from the following model:

$$Y = X_3^3 + X_3^2 + X_3 + \frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6 + 2I(X_{p_1+1} = 0) + I(X_{p_1+1} = 1) + \epsilon.$$

We consider different distributions for ϵ: [(1) standard normal] $\epsilon \sim N(0, 1)$; [(2) heteroscedastic] $\epsilon = (0.5+\zeta^\top \mathbf{x})\epsilon^*$, where $\zeta = (0, 1, 0, ..., 0) \in \mathbb{R}^{3p_1+2(p-p_1)}$, $\epsilon^* \sim N(0, 1)$; and [(3) asymmetric and heteroscedastic] $\epsilon = (0.5 + \zeta^\top \mathbf{x})\epsilon^{**}$, where $\epsilon^{**} \sim \chi^2(3)$. These simulations will verify our theoretical results that the conditional quantile estimates remain consistent even in the presence of heteroscedastic errors, an often cited strength of quantile regression compared to mean regression. The setting with asymmetric errors is of particular interest when modeling the median, because in that case the Huber estimator will provide a biased estimate of the conditional median. When $\tau \neq .5$, then the Huber estimator will always have some bias, hence the need for a sufficiently small γ. In our simulations, we set $p = 100$ and $p_1 = 60$. Linear, quadratic and cubic terms are included for the first p_1 covariates and two dummy variables are used for each of the remaining $p - p_1$ variables, so that the actual dimension of the covariates is 260. We generate the data with different sample sizes of $n \in \{100, 200, 500\}$, and for each setting, we conduct 100 replicates.

We compare the proposed Huber quantile regression group lasso estimator (hrq-glasso) with the following alternative approaches:

1. SOCP: solving for quantile loss with the group lasso penalty via a direct implementation of second order cone programming with MOSEK;
2. grpreg: the least squares group-lasso using the R package `grpreg` (Breheny and Zeng, 2017) which implements the algorithm proposed by Breheny and Huang (2015);

3. hqreg: the semismooth Newton method (Yi and Huang, 2017) for Huber lasso using the R package `hqreg` (Yi, 2017);
4. glmnet: the least squares lasso (Tibshirani, 1996) using the R package `glmnet` (Friedman et al, 2010).

We fit the above models for $\tau \in \{.1, .3, .5, .7, .9\}$. The least squares group-lasso and lasso do not directly model a conditional quantile. For these methods we use a naive method of shifting the estimated intercept by $\hat{\sigma}\Phi^{-1}(\tau)$ in order to approximate the τth quantile of predicted value, where $\hat{\sigma}^2$ is estimated by the variance of the residuals. This is a naive solution that will fail to provide consistent estimates if the distribution is not normal or the variance is not constant.

The compared measures include CPU time, model error (ME) that is defined as

$$\text{ME}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}(\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda} - \boldsymbol{\beta}^{*}_{\tau})^{\top} X^{\top} X (\hat{\boldsymbol{\beta}}^{\tau}_{\gamma,\lambda} - \boldsymbol{\beta}^{*}_{\tau}), \tag{11}$$

the number of false positives, $\sum_{j=1}^{p} I(\beta^{*}_{\tau,j} = 0, \hat{\beta}_j \neq 0)$, and false negatives, $\sum_{j=1}^{p} I(\beta^{*}_{\tau,j} \neq 0, \hat{\beta}_j = 0)$. As noted in Fan and Li (2001), the size of model error reflects model selection performance, which accounts for prediction error. Thus to make a fair comparison of the best performance of each approach, the tuning parameter λ is selected to minimize ME for each method in each replicate. For our method, the sequence of λ values and the values for γ are selected as described in the previous section. The same sequence of λ is supplied to SOCP as it is the closest competitor to our approach.

Figure 2 depicts the computing time (in logarithm) of hrq-glasso and SOCP implemented with MOSEK. Other approaches are excluded because they are optimizing different objective functions. The proposed approach is consistently faster than the exact SOCP across all settings, and the deviation grows with sample size n increasing. All programs ran on a Linux server with Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz and 64GB RAM.

Tables 1 – 3 provide the average ME (and standard deviations) of the five different approaches for the three different error distributions. In all three tables the group lasso methods outperform the lasso counterparts verifying the importance of incorporating group structure into the penalty. As expected, the least squares methods perform best when the errors are normally distributed and the quantile approaches do better when the error has a heteroscedastic or asymmetric distribution. The proposed approach, which approximates the quantile loss, performs competitively with the exact SOCP algorithm. In fact, in some settings the Huber approximation (hrq-glasso) performs better than the exact approach (SOCP). To briefly summarize, the group lasso methods provide more accurate estimates, the quantile methods perform better than the mean regression methods in the non-normal error settings and the performance of SOCP and hrq-glasso are similar.

Tables 4 – 6 provide model selection results for the three different error settings. Across the three settings there are some consistent findings. First, the group lasso methods tend to pick larger models resulting in more false positives,

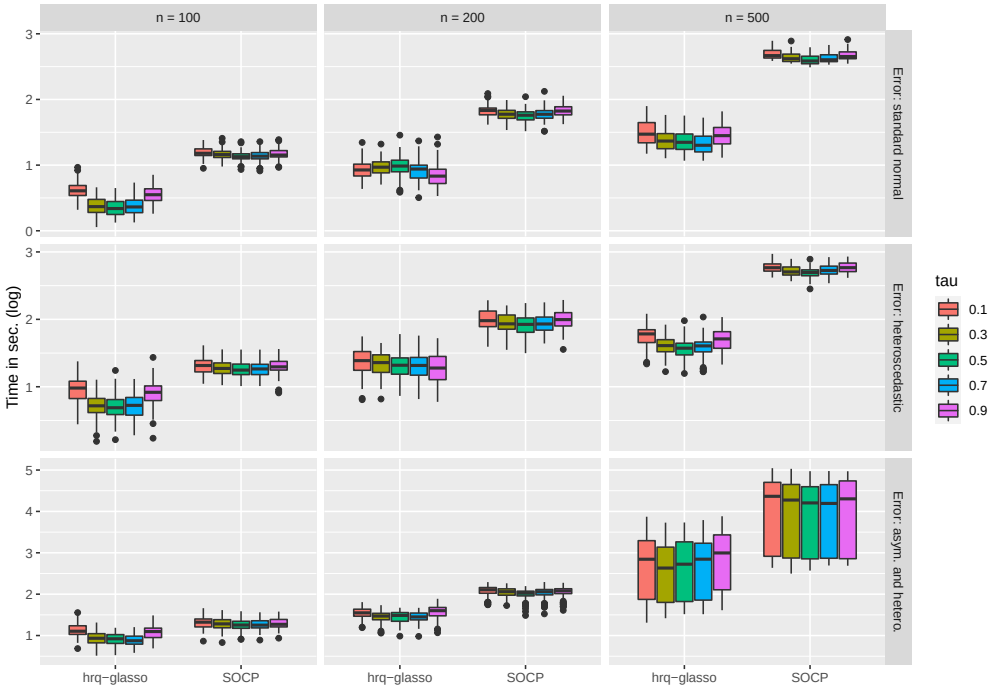

Fig. 2 Comparison of CPU time (in logarithm scale) between the proposed approach and the SOCP with MOSEK. Rows are for different types of error and columns are for different sample size.

but fewer false negatives. The least squares methods perform better when the errors are normally distributed and worse for the noisier error distributions. One noticeable exception, seen in Table 6, is that for setting three grpreg is competitive with respect to false negatives and tends to have fewer false positive. However, as n increases we can see the quantile methods tend to have fewer false negatives. Performance between hrq-glasso and SOCP are very similar, suggesting that any bias caused by using the Huber approximation has a small impact on model selection.

In summary, the proposed hrq-glasso produces the same accuracy as the benchmark method SOCP, which directly solves the optimization problems using the modern software MOSEK. However, the hrq-glasso gains significant computational efficiency with little sacrifice of accuracy. Although the existing quantile lasso methods deliver reasonably good estimates, they are not ideal choices when there is a group structure to the predictors.

In the results discussed so far γ is selected as outlined in the previous section. In addition, different models of hrq-glasso were fit for different values of γ. This was done to examine how sensitive the performance is to the choice of γ. Figure 3 shows performance of our algorithm across different values of γ in terms of the model error, as defined in (11), for $\tau = 0.3$ (left), 0.5 (middle)

Table 1 Comparison of model error (ME) between hrq-glasso, SOCP, hqreg and grpreg for quantile regression under standard normal error.

N	τ	hrq-glasso	SOCP	hqreg	grpreg	glmnet
100	0.1	0.934 (0.336)	1.061 (0.374)	1.378 (0.459)	0.332 (0.133)	0.429 (0.120)
	0.3	0.500 (0.183)	0.543 (0.197)	0.661 (0.192)	0.322 (0.132)	0.395 (0.111)
	0.5	0.415 (0.138)	0.444 (0.138)	0.550 (0.150)	0.320 (0.134)	0.380 (0.108)
	0.7	0.496 (0.169)	0.530 (0.177)	0.659 (0.195)	0.323 (0.138)	0.388 (0.113)
	0.9	0.910 (0.313)	1.032 (0.330)	1.325 (0.423)	0.335 (0.147)	0.418 (0.123)
200	0.1	0.481 (0.181)	0.531 (0.180)	0.670 (0.182)	0.248 (0.101)	0.208 (0.062)
	0.3	0.259 (0.095)	0.274 (0.097)	0.336 (0.102)	0.240 (0.095)	0.199 (0.059)
	0.5	0.228 (0.081)	0.241 (0.087)	0.301 (0.089)	0.238 (0.094)	0.197 (0.058)
	0.7	0.255 (0.094)	0.272 (0.099)	0.355 (0.106)	0.240 (0.096)	0.200 (0.059)
	0.9	0.485 (0.188)	0.542 (0.174)	0.736 (0.208)	0.249 (0.103)	0.210 (0.062)
500	0.1	0.179 (0.071)	0.196 (0.072)	0.220 (0.072)	0.055 (0.020)	0.077 (0.023)
	0.3	0.098 (0.032)	0.103 (0.035)	0.122 (0.034)	0.053 (0.019)	0.074 (0.022)
	0.5	0.087 (0.026)	0.090 (0.028)	0.110 (0.027)	0.053 (0.019)	0.074 (0.022)
	0.7	0.097 (0.032)	0.102 (0.035)	0.125 (0.033)	0.053 (0.020)	0.074 (0.023)
	0.9	0.173 (0.063)	0.190 (0.065)	0.231 (0.078)	0.056 (0.020)	0.077 (0.024)

Table 2 Comparison of model error (ME) between hrq-glasso, SOCP, hqreg and grpreg for quantile regression under heteroscedastic error.

N	τ	hrq-glasso	SOCP	hqreg	grpreg	glmnet
100	0.1	4.486 (1.912)	4.606 (1.967)	5.372 (2.213)	5.038 (2.523)	5.357 (2.604)
	0.3	1.586 (0.657)	1.620 (0.648)	2.117 (0.942)	2.125 (1.242)	2.469 (1.304)
	0.5	0.993 (0.645)	1.003 (0.650)	1.390 (0.691)	1.499 (0.962)	1.873 (1.101)
	0.7	1.545 (0.814)	1.574 (0.816)	2.034 (0.909)	2.017 (1.242)	2.414 (1.477)
	0.9	4.249 (1.824)	4.355 (1.836)	4.954 (1.956)	4.824 (2.550)	5.257 (2.874)
200	0.1	2.752 (1.439)	2.769 (1.303)	3.096 (1.489)	4.422 (1.463)	4.649 (1.599)
	0.3	0.937 (0.441)	0.936 (0.407)	1.083 (0.440)	1.521 (0.689)	1.780 (0.744)
	0.5	0.438 (0.268)	0.440 (0.262)	0.565 (0.294)	0.901 (0.537)	1.158 (0.578)
	0.7	0.786 (0.410)	0.798 (0.408)	0.962 (0.441)	1.397 (0.781)	1.636 (0.894)
	0.9	2.195 (1.360)	2.253 (1.354)	2.576 (1.396)	4.086 (1.799)	4.252 (1.999)
500	0.1	0.879 (0.457)	0.884 (0.446)	0.808 (0.415)	4.023 (1.000)	4.130 (1.052)
	0.3	0.366 (0.186)	0.367 (0.183)	0.344 (0.173)	1.079 (0.400)	1.197 (0.420)
	0.5	0.151 (0.065)	0.153 (0.064)	0.176 (0.062)	0.471 (0.285)	0.602 (0.279)
	0.7	0.385 (0.181)	0.388 (0.181)	0.372 (0.189)	1.065 (0.428)	1.202 (0.459)
	0.9	0.864 (0.411)	0.877 (0.413)	0.827 (0.407)	4.000 (1.065)	4.125 (1.145)

and 0.7 (right). The results are based on 500 replicates under the same error settings with $n = 200$. From all cases shown in the figure, it can be clearly seen that the model errors are close for a wide range of γ, e.g., $0.05 \leq \gamma \leq 2$, while for both small and large extreme γ the model errors become larger. This experiment implies that our algorithm is relatively insensitive to the choice of γ within a proper range.

5 Application to Ames Housing Data

In this section, we demonstrate the utility of the proposed estimator empirically by applying it to a real dataset – the Ames Housing data (De Cock,

Table 3 Comparison of model error (ME) between hrq-glasso, SOCP, hqreg and grpreg for quantile regression under heteroscedastic and asymmetric error.

N	τ	hrq-glasso	SOCP	hqreg	grpreg	glmnet
100	0.1	8.441 (4.116)	8.505 (4.119)	8.959 (4.633)	20.998 (15.107)	20.816 (14.155)
	0.3	5.430 (2.619)	5.437 (2.624)	6.166 (3.106)	11.971 (8.342)	11.631 (7.161)
	0.5	3.909 (2.135)	3.909 (2.115)	4.716 (2.456)	8.937 (7.161)	8.849 (6.035)
	0.7	5.232 (2.541)	5.258 (2.498)	5.956 (2.818)	9.939 (8.143)	10.161 (7.075)
	0.9	22.368 (9.102)	22.395 (9.093)	22.992 (9.463)	26.920 (11.692)	27.336 (11.417)
200	0.1	3.116 (1.908)	3.151 (1.910)	3.471 (1.929)	19.896 (10.862)	20.097 (10.651)
	0.3	2.685 (1.700)	2.683 (1.697)	2.768 (1.788)	9.892 (5.964)	9.811 (5.086)
	0.5	2.139 (1.188)	2.166 (1.135)	2.636 (1.271)	6.669 (4.669)	6.887 (3.886)
	0.7	3.489 (1.455)	3.493 (1.429)	4.198 (1.633)	7.744 (4.588)	8.398 (4.224)
	0.9	17.805 (6.913)	17.701 (6.858)	17.380 (7.559)	25.629 (7.536)	26.642 (7.555)
500	0.1	0.986 (0.658)	0.994 (0.652)	0.804 (0.539)	18.618 (7.136)	18.995 (7.291)
	0.3	1.020 (0.716)	0.998 (0.683)	0.868 (0.611)	7.463 (3.409)	7.726 (3.248)
	0.5	1.098 (0.549)	1.081 (0.541)	1.110 (0.576)	4.265 (2.473)	4.641 (2.349)
	0.7	1.829 (0.583)	1.820 (0.566)	2.057 (0.668)	5.605 (2.328)	6.093 (2.343)
	0.9	9.698 (4.690)	9.633 (4.668)	8.849 (4.489)	24.095 (4.836)	24.763 (5.000)

Table 4 Comparison of false positive (FP) and false negative (FN) between hrq-glasso, SOCP, hqreg and grpreg for quantile regression under the setting of standard normal error.

N	τ	hrq-glasso	SOCP	hqreg	grpreg	glmnet
			False positive (FP)			
100	0.1	68.83 (22.19)	66.64 (20.27)	54.01 (15.71)	21.53 (11.2)	22.62 (5.43)
	0.3	71.95 (24.80)	70.57 (25.66)	52.45 (17.91)	22.74 (12.07)	27.55 (6.05)
	0.5	68.37 (22.81)	68.03 (21.46)	48.10 (17.93)	22.89 (12.2)	30.03 (6.59)
	0.7	73.15 (24.09)	71.17 (22.84)	52.37 (16.17)	22.71 (11.98)	27.93 (6.17)
	0.9	71.27 (20.07)	66.51 (20.42)	53.36 (13.41)	21.70 (11.04)	23.18 (5.75)
200	0.1	79.05 (27.90)	75.70 (23.67)	59.53 (15.96)	10.67 (8.08)	25.27 (6.27)
	0.3	69.94 (21.83)	68.47 (21.75)	47.58 (13.38)	10.75 (8.17)	28.76 (6.05)
	0.5	67.93 (20.17)	65.31 (22.77)	46.10 (15.73)	10.80 (8.27)	29.69 (5.93)
	0.7	64.89 (23.11)	63.31 (22.33)	50.04 (17.15)	10.80 (8.27)	28.53 (5.58)
	0.9	73.58 (23.55)	73.18 (20.99)	60.61 (15.67)	10.77 (8.19)	24.85 (5.66)
500	0.1	67.22 (25.87)	65.81 (26.14)	38.11 (15.03)	36.13 (8.92)	25.97 (5.54)
	0.3	61.52 (18.87)	58.97 (23.06)	34.47 (12.42)	38.86 (9.27)	28.19 (5.78)
	0.5	59.39 (18.00)	55.86 (17.72)	32.09 (11.04)	40.37 (9.63)	29.02 (5.96)
	0.7	58.61 (17.71)	56.60 (21.73)	33.12 (11.48)	39.90 (9.83)	28.34 (5.67)
	0.9	71.30 (25.62)	70.09 (23.79)	42.80 (13.76)	37.92 (12.23)	25.95 (5.64)
			False negative (FN)			
100	0.1	0.00 (0.00)	0.00 (0.00)	0.50 (0.67)	0.00 (0.00)	0.33 (0.51)
	0.3	0.00 (0.00)	0.00 (0.00)	0.34 (0.52)	0.00 (0.00)	0.19 (0.39)
	0.5	0.00 (0.00)	0.00 (0.00)	0.31 (0.46)	0.00 (0.00)	0.17 (0.38)
	0.7	0.00 (0.00)	0.00 (0.00)	0.37 (0.51)	0.00 (0.00)	0.22 (0.42)
	0.9	0.02 (0.20)	0.02 (0.20)	0.64 (0.67)	0.00 (0.00)	0.31 (0.51)
200	0.1	0.00 (0.00)	0.00 (0.00)	0.07 (0.26)	0.00 (0.00)	0.00 (0.00)
	0.3	0.00 (0.00)	0.00 (0.00)	0.03 (0.17)	0.00 (0.00)	0.00 (0.00)
	0.5	0.00 (0.00)	0.00 (0.00)	0.03 (0.17)	0.00 (0.00)	0.00 (0.00)
	0.7	0.00 (0.00)	0.00 (0.00)	0.01 (0.10)	0.00 (0.00)	0.00 (0.00)
	0.9	0.00 (0.00)	0.00 (0.00)	0.10 (0.30)	0.00 (0.00)	0.01 (0.10)
500	0.1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.3	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.7	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.9	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Table 5 Comparison of false positive (FP) and false negative (FN) between hrq-glasso, SOCP, hqreg and grpreg for quantile regression under the setting of heteroscedastic error.

N	τ	hrq-glasso	SOCP	hqreg	grpreg	glmnet
			False positive (FP)			
100	0.1	72.81 (34.60)	66.26 (22.43)	45.86 (17.31)	35.27 (15.40)	21.98 (9.55)
	0.3	68.23 (25.58)	66.43 (24.59)	48.02 (18.62)	33.73 (12.00)	22.11 (8.68)
	0.5	60.97 (25.59)	60.37 (22.99)	41.85 (15.31)	33.51 (11.37)	22.78 (7.35)
	0.7	66.32 (24.76)	64.40 (23.45)	46.87 (16.16)	33.39 (12.70)	23.07 (7.42)
	0.9	64.69 (22.85)	60.49 (21.82)	41.52 (19.21)	34.70 (15.02)	22.95 (8.31)
200	0.1	84.39 (32.29)	82.93 (26.92)	59.34 (18.28)	35.29 (16.14)	26.82 (12.29)
	0.3	74.06 (31.40)	71.69 (29.49)	48.30 (19.46)	30.68 (12.24)	22.71 (8.12)
	0.5	63.03 (22.41)	61.68 (23.19)	38.20 (13.66)	29.91 (10.16)	22.85 (6.20)
	0.7	68.74 (23.85)	68.44 (25.55)	49.55 (19.11)	33.01 (14.02)	25.64 (8.68)
	0.9	80.86 (24.37)	79.41 (23.89)	60.67 (18.64)	39.66 (20.09)	32.44 (13.04)
500	0.1	90.09 (31.10)	88.92 (29.74)	49.30 (19.99)	67.24 (38.35)	39.74 (19.99)
	0.3	89.55 (28.26)	90.06 (34.28)	46.63 (19.31)	38.44 (17.83)	26.08 (9.38)
	0.5	61.26 (19.83)	60.18 (23.44)	29.45 (11.36)	31.97 (10.77)	24.34 (6.56)
	0.7	86.12 (27.01)	87.76 (30.50)	44.71 (18.04)	39.93 (20.20)	28.45 (9.78)
	0.9	93.68 (30.01)	92.80 (29.44)	51.81 (18.87)	69.30 (42.90)	44.36 (19.65)
			False negative (FN)			
100	0.1	0.39 (0.58)	0.39 (0.53)	1.78 (1.18)	0.71 (0.70)	2.19 (1.08)
	0.3	0.54 (0.50)	0.52 (0.50)	1.33 (0.95)	0.78 (0.84)	2.26 (1.23)
	0.5	0.02 (0.20)	0.02 (0.20)	0.70 (0.66)	0.24 (0.65)	1.52 (1.18)
	0.7	0.54 (0.50)	0.57 (0.56)	1.67 (1.06)	0.86 (0.89)	2.19 (1.20)
	0.9	0.70 (1.16)	0.77 (1.30)	2.44 (1.48)	0.82 (0.83)	2.26 (1.18)
200	0.1	0.05 (0.22)	0.07 (0.26)	0.37 (0.66)	0.36 (0.52)	1.00 (0.72)
	0.3	0.41 (0.49)	0.44 (0.50)	0.50 (0.52)	0.36 (0.50)	1.06 (0.75)
	0.5	0.00 (0.00)	0.00 (0.00)	0.06 (0.28)	0.00 (0.00)	0.53 (0.56)
	0.7	0.26 (0.44)	0.25 (0.44)	0.35 (0.58)	0.34 (0.50)	0.97 (0.80)
	0.9	0.02 (0.14)	0.03 (0.17)	0.53 (0.82)	0.31 (0.46)	0.83 (0.75)
500	0.1	0.00 (0.00)	0.00 (0.00)	0.01 (0.10)	0.28 (0.45)	0.58 (0.59)
	0.3	0.02 (0.14)	0.02 (0.14)	0.02 (0.14)	0.38 (0.49)	0.64 (0.59)
	0.5	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.12 (0.33)
	0.7	0.01 (0.10)	0.01 (0.10)	0.00 (0.00)	0.36 (0.48)	0.60 (0.57)
	0.9	0.00 (0.00)	0.00 (0.00)	0.02 (0.14)	0.24 (0.43)	0.57 (0.56)

2011; Kuhn, 2020). The dataset contains detailed information of 2930 residential properties sold between 2006 and 2010 in Ames, Iowa. There are 82 variables including the observation number, which will be excluded from the analysis, and the sale price, which is the response variable in our analysis. The remaining 80 explanatory variables, measured in nominal, ordinal and continuous scales, provide detailed information about each individual property. More details about the Ames data can be found in De Cock (2011).

Among the 80 explanatory variable, there are 37 nominal and 23 ordinal variables, which are typically converted to dummy variables in regression analysis. In this case, group-lasso-type methods are desirable if the number of variables is large and selecting important variables are of interest. In our study, we apply the proposed method to select important explanatory variables and build quantile regression models to predict the sale price. Our exploratory data analysis suggests that data pre-processing is needed before fitting the regression models. In particular, we first recode several nominal variables with

Table 6 Comparison of false positive (FP) and false negative (FN) between hrq-glasso, SOCP, hqreg and grpreg for quantile regression under the setting of heteroscedastic and asymmetric error.

N	τ	hrq-glasso	SOCP	hqreg	grpreg	glmnet
			False positive (FP)			
100	0.1	58.90 (16.27)	55.68 (16.04)	33.38 (13.35)	42.11 (16.45)	21.79 (10.09)
	0.3	60.01 (17.64)	56.98 (19.74)	34.36 (14.26)	26.63 (14.43)	12.33 (7.26)
	0.5	51.70 (20.33)	49.13 (18.60)	29.47 (12.92)	25.94 (12.78)	12.03 (7.00)
	0.7	42.92 (17.31)	39.79 (16.94)	21.83 (12.98)	30.64 (12.43)	14.29 (6.76)
	0.9	27.10 (16.69)	23.50 (14.49)	12.80 (10.07)	25.78 (15.71)	12.48 (8.80)
200	0.1	68.42 (17.65)	69.29 (17.13)	50.66 (15.90)	67.55 (26.53)	37.50 (15.48)
	0.3	73.98 (22.68)	74.57 (24.60)	43.42 (17.97)	29.09 (16.59)	15.13 (9.09)
	0.5	59.62 (25.72)	57.91 (25.24)	35.62 (17.95)	25.35 (12.03)	13.74 (7.64)
	0.7	51.05 (25.53)	49.20 (23.66)	30.74 (19.14)	33.72 (16.38)	18.44 (9.30)
	0.9	41.44 (24.81)	36.76 (21.21)	18.71 (13.86)	35.77 (26.76)	20.32 (14.93)
500	0.1	86.89 (23.48)	86.69 (22.46)	46.57 (18.30)	114.48 (38.92)	61.32 (20.67)
	0.3	88.10 (24.26)	88.54 (27.46)	42.73 (14.86)	39.52 (20.98)	20.22 (9.49)
	0.5	81.65 (29.02)	80.52 (31.86)	40.72 (19.56)	32.81 (14.71)	19.18 (7.04)
	0.7	65.23 (27.16)	65.28 (28.29)	35.47 (15.30)	43.77 (17.72)	27.20 (8.89)
	0.9	74.27 (30.46)	69.97 (27.26)	32.00 (14.65)	58.64 (43.44)	33.84 (23.82)
			False negative (FN)			
100	0.1	0.60 (1.15)	0.63 (1.14)	2.85 (1.38)	1.71 (1.65)	4.30 (1.73)
	0.3	0.70 (1.08)	0.78 (1.25)	2.71 (1.39)	2.59 (2.28)	4.92 (1.74)
	0.5	1.22 (1.25)	1.27 (1.28)	3.27 (1.55)	2.61 (2.25)	4.97 (1.77)
	0.7	2.24 (1.76)	2.27 (1.75)	4.40 (1.44)	2.03 (1.71)	4.72 (1.52)
	0.9	4.24 (1.93)	4.36 (1.95)	6.06 (1.32)	2.20 (1.44)	5.02 (1.32)
200	0.1	0.02 (0.20)	0.02 (0.20)	0.52 (0.70)	0.69 (0.98)	2.78 (1.48)
	0.3	0.06 (0.28)	0.06 (0.28)	0.57 (0.73)	1.61 (1.46)	3.88 (1.78)
	0.5	0.40 (0.51)	0.41 (0.51)	1.42 (1.10)	1.59 (1.39)	3.93 (1.70)
	0.7	1.04 (1.19)	1.06 (1.18)	2.84 (1.28)	1.28 (1.26)	3.52 (1.59)
	0.9	2.94 (2.01)	2.89 (1.92)	4.65 (1.31)	1.55 (1.25)	3.67 (1.51)
500	0.1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.21 (0.43)	1.40 (1.11)
	0.3	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.83 (1.06)	2.14 (1.30)
	0.5	0.19 (0.39)	0.20 (0.40)	0.27 (0.55)	0.85 (1.01)	2.11 (1.26)
	0.7	0.46 (0.50)	0.48 (0.50)	1.04 (0.86)	0.58 (0.78)	1.78 (0.91)
	0.9	0.43 (1.01)	0.38 (1.02)	2.46 (1.19)	0.74 (0.94)	1.98 (1.07)

binary values due to the high frequency for one category and extremely low frequencies for all other categories. For example, the variable "Pool_QC" (pool quality) has five categories: No_pool (2,917), Typical (3), Fair (2), Good (4), Excellent (4), where their frequencies are in the parenthesis. We recode this variable using binary values to indicate if there is a pool in the property. We also convert some continuous variables to nominal such as "Mo_Sold" (month sold) as the numerical value of month does not have quantitative meaning. To further avoid rare categories, we filter out all observations such that for any categorical variables, the remaining data has at least 10 observations for each category. Finally, the categorical variables with a single category are removed. The supplemental material includes code with all of the pre-processing steps. The final data has 2,083 observations and 72 explanatory variables, which due to the presence of categorical predictors results in a dimension of 201 for the design matrix.

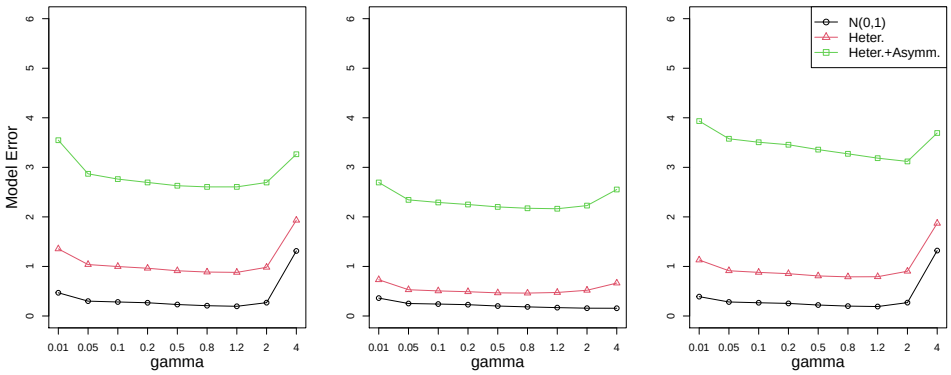

Fig. 3 Comparison of model error (defined in (11)) across different values of tuning parameter γ for quantile regression models with $\tau = 0.3$ (left), $\tau = 0.5$ (middle) and $\tau = 0.7$ (right), under different simulation settings.

To quantify the prediction accuracy, we randomly select 1,600 observations as training set and leave the remaining as the testing set, and repeat this process 100 times. In fitting each quantile regression ($\tau = 0.1$, 0.3, 0.5, 0.7, and 0.9), a 5-fold cross-validation is applied to select the tuning parameter λ. The quantile check loss is used for cross-validation. The five approaches used in the simulation studies are applied here and compared via several metrics, which are elaborated next.

We first compare the computational time between hrq-lasso and SOCP as shown in Figure 4. Similar to the simulation studies, we exclude the other three approaches for fair comparison. The direct implementation with SOCP is noticeably slower than hrq-lasso across all five quantiles. This is expected when the sample size is large, which is also demonstrated in the simulation studies.

To asses the prediction accuracy at quantiles we consider two closely related metrics: (1) out-of-sample check loss and (2) overall predicted quantile. The first metric is natural and straightforward to calculate, while the latter is defined as $\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(Y_i < \hat{Y}_i^{\tau})$, where n is the testing sample size. A consistent estimator of the τth quantile will provide a value that is close to τ. The left panel of Figure 5 shows the out-of-sample check loss and right panel shows the overall predicted quantiles for all five approaches. As expected, hrq-glasso and SOCP performs quite similarly in terms of the prediction accuracy as both approaches are essentially solving the same optimization problem, while hrq-glasso delivers slightly larger loss at the 10th and 90th percentile. The naive approach using group lasso performs well in terms of out-of-sample check loss, but its overall predicted quantiles are highly biased.

Lastly, we show the model size in Table 7. Reported are average of the number of nonzero coefficients and the standard deviation over 100 replications. The results for hrq-glasso and SOCP are again very close as expected,

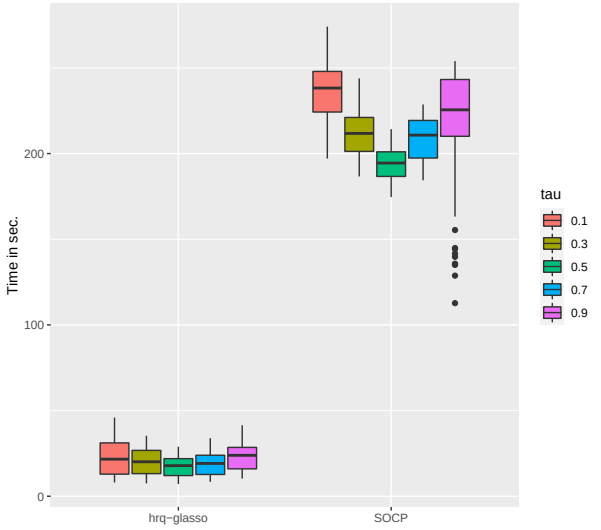

Fig. 4 Comparison of computing time between hrq-glasso and SOCP applied to the Ames Housing data.

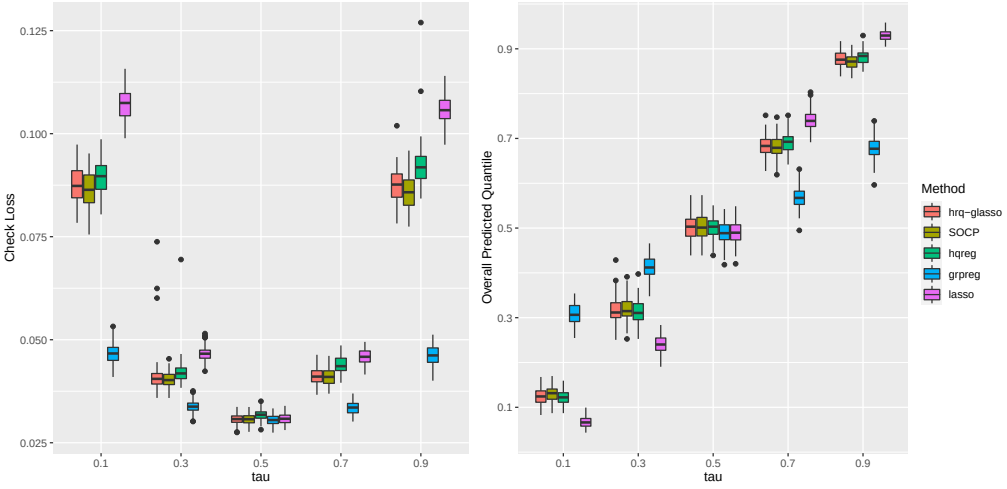

Fig. 5 Comparison of out-of-sample check loss and overall predicted quantiles across different methods applied to the Ames Housing data.

while hrq-glasso has slightly higher variance at τ =0.1 and 0.9. Their average number of nonzero coefficients are also close to the naive quantile group lasso approach (grpreg). The quantile lasso (hqreg) and naive approach with lasso (glmnet) tend to fit smaller models. This is also expected because both of them use L_1 penalty which does not enforce grouped variable selection.

Table 7 Average number of nonzero coefficients based on different methods. In parenthesis are standard deviations.

τ	hrq-glasso	SOCP	hqreg	grpreg	glmnet
0.1	173.81 (14.31)	174.05 (13.36)	93.41 (6.64)	189.72 (4.61)	138.30 (10.80)
0.3	192.06 (3.52)	189.61 (4.08)	95.12 (4.56)	189.72 (4.61)	138.30 (10.80)
0.5	186.76 (5.50)	184.61 (5.91)	90.09 (3.68)	189.72 (4.61)	138.30 (10.80)
0.7	189.28 (4.72)	189.57 (4.35)	91.50 (4.07)	189.72 (4.61)	138.30 (10.80)
0.9	174.34 (12.87)	177.93 (10.56)	82.94 (6.06)	189.72 (4.61)	138.30 (10.80)

In summary, our proposed method demonstrates high utility in terms of both computing time and prediction accuracy through a real data example. Note that SOCP is the only competitor among all alternative approaches to achieve the goal of group-wise variable selection and estimating quantile regression. Comparing to SOCP, the proposed method is much faster, while accuracy results are very similar.

6 Conclusion

We propose using a Huber approximation of quantile loss and a group lasso penalty to simultaneously perform model selection and estimation for conditional quantile models. A groupwise-majorization descent algorithm is used, which cannot be applied to the non-differentiable quantile loss. This provides significant computational gains. In addition, we demonstrate that rates of convergence for this approach is the same as if the quantile loss was used. The implementation of the proposed approach is publicly available on CRAN(Li and Sherwood, 2021; Sherwood et al, 2022).

7 Supplemental

Supplemental material contains proofs for the theoretical results and technical derivations for the algorithm.

References

Alfo M, Salvati N, Ranallli MG (2017) Finite mixtures of quantile and m-quantile regression models. Stat Comput 27(2):547–570

Belloni A, Chernozhukov V (2011) L1-penalized quantile regression in high-dimensional sparse models. Ann Statist 39(1):82–130

Bianchi A, Salvati N (2015) Asymptotic properties and variance estimators of the m-quantile regression coefficients estimators. Commun Stat Theory Methods 44(11):2416–2429

Bianchi A, Fabrizi E, Salvati N, et al (2018) Estimation and testing in m-quantile regression with applications to small area estimation. Int Stat Rev 86(3):541–570

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of lasso and dantzig selector. Ann Statist 37(4):1705–1732

Breckling J, Chambers R (1988) M-quantiles. Biometrika 75(4):761–771

Breheny P, Huang J (2015) Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat Comput 25:173–187

Breheny P, Zeng Y (2017) grpreg: Regularization Paths for Regression Models with Grouped Covariates 3.1-2. R package version 3.3.1

Chambers R, Tzavidis N (2006) M-quantile models for small area estimation. Biometrika 93(2):255–268

Ciuperca G (2019) Adaptive group lasso selection in quantile models. Statist Papers 60(1):173–197

De Cock D (2011) Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. J Stat Educ 19(3):1–15

Del Sarto S, Marino MF, Ranalli MG, et al (2019) Using finite mixtures of m-quantile regression models to handle unobserved heterogeneity in assessing the effect of meteorology and traffic on air quality. Stoch Environ Res Risk Assess 33(7):1345–1359

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc 96(456):1348–1360

Fan J, Li Q, Wang Y (2017) Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. J R Stat Soc Ser B Stat Methodol 79(1):247–265

Fasiolo M, Wood SN, Zaffran M, et al (2021) Fast calibrated additive quantile regression. J Amer Statist Assoc 116(535):1402–1412

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22

He X, Shao QM (2000) On parameters of increasing dimensions. J Multivariate Anal 73:120–135

Huang J, Zhang T (2010) The benefit of group sparsity. Ann Statist 38(4):1978–2004

Huang J, Horowitz JL, Wei F (2010) Variable selection in nonparametric additive models. Ann Statist 38(4):2282–2313

Huber PJ (1964) Robust estimation of a location parameter. Ann Math Statist 35(1):73–101

Huber PJ (1973) Robust regression: Asymptotics, conjectures and monte carlo. Ann Statist 1(5):799–821

Hunter DR, Lange K (2004) A tutorial on mm algorithms. Amer Statist 58(1):30–37

Kato K (2011) Group lasso for high dimensional sparse quantile regression models. URL https://arxiv.org/abs/1103.1458

Koenker R (2005) Quantile Regression. Cambridge University Press

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46(1):33–50

Koenker R, Mizera I (2014) Convex optimization in R. J Stat Softw 60(5):1–23

Koenker RW, D'Orey V (1987) Computing regression quantiles. J R Stat Soc Ser C Appl Stat 36(3):383–393

Kokic P, Chambers R, Breckling J, et al (1997) A measure of production performance. J Bus Econ Stat 15(4):445–451

Kuhn M (2020) AmesHousing: The Ames Iowa Housing Data. R package version 0.0.4

Lee Y, MacEachern SN, Jung Y (2012) Regularization of case-specific parameters for robustness and efficiency. Statist Sci 27(3):350–372

Li S, Sherwood B (2021) hrqglas: Group Variable Selection for Quantile and Robust Mean Regression. R package version 1.0.1

Liu LZ, Wu FX, Zhang WJ (2014) A group lasso-based method for robustly inferring gene regulatory networks from multiple time-course datasets. BMC Syst Biol 8:1–12

Lounici K, et al (2011) Oracle inequalities and optimal inference under group sparsity. Ann Statist 39(4):2164–2204

Meier L, Van De Geer S, Bühlmann P (2008) The group lasso for logistic regression. J R Stat Soc Ser B Stat Methodol 70(1):53–71

Muggeo VM, Sciandra M, Augugliaro L (2012) Quantile regression via iterative least squares computations. J Stat Comput Simul 82(11):1557–1569

Negahban SN, et al (2012) A unified framework for high-dimensional analysis fo m-estimators with decomposable regualrizers. Statist Sci 27(4):538–557

Newey WK, Powell JL (1987) Asymmetric least squares estimation and testing. Econometrica 55(4):819–847

Portnoy S, Koenker R (1997) The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. Statist Sci 12(4):279–300

Pratesi M, Ranalli MG, Salvati N (2009) Nonparametric m-quantile regression using penalised splines. J Nonparametr Stat 21(3):287–304

Sherwood B, Maidman A, Li S (2022) rqPen: Penalized Quantile Regression. R package version 3.0

Sun Q, Zhou WX, Fan J (2020) Adaptive huber regression. J Amer Statist Assoc 115(529):254–265

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 58(1):267–288

Tibshirani R, et al (2012) Strong rules for discarding predictors in lasso-type problems. J R Stat Soc Ser B Stat Methodol 74(2):245–266

Tzavidis N, Salvati N, Schmid T, et al (2016) Longitudinal analysis of the strengths and difficulties questionnaire scores of the millennium cohort study children in england using m-quantile random-effects regression. J R Stat Soc Ser A Stat Soc 179(2):427–452

Wang L, Wu Y, Li R (2012) Quantile regression of analyzing heterogeneity in ultra-high dimension. J Amer Statist Assoc 107(497):214–222

Wu TT, Lange K, et al (2010) The MM alternative to EM. Statist Sci 25(4):492–505

Wu Y, Liu Y (2009) Variable selection in quantile regression. Statist Sinica 19(2):801–817

Xu J, Ying Z (2010) Simultaneous estimation and variable selection in median regression using lasso-type penalty. Ann Inst Statist Math 62:487–514

Yang Y, Zou H (2015) A fast unified algorithm for solving group-lasso penalize learning problems. Stat Comput 25(6):1129–1141

Yi C (2017) hqreg: Regularization Paths for Lasso or Elastic-Net Penalized Huber Loss Regression and Quantile Regression. R package version 1.4

Yi C, Huang J (2017) Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. J Comput Graph Statist 26(3):547–557

Yuan M, Lin Y (2005) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol 68(1):49–67

Zhou WX, et al (2018) A new perspective on robust m-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. Ann Statist 46(5):1904–1931

Supplemental Material to Quantile regression feature selection and estimation with grouped variables using Huber approximation

Proofs

Throughout the proofs C will represent a generic constant whose value may change from line to line.

Proof of Theorem 1

Proof. When Conditions 1 and 2 hold the standard quantile regression estimator, $\hat{\boldsymbol{\beta}}^{\tau}_{\gamma=0}$, has the same limiting distribution, see Theorem 4.1 from Koenker (2005).

Note that,

$$\left|h^{\tau}_{\gamma}(u) - 2\rho_{\tau}(u)\right| = |h_{\gamma}(u) - |u|| \leq \gamma/2. \tag{1}$$

Using the above inequality and that $\gamma = o(1)$ it follows that

$$\frac{1}{2n}\sum_{i=1}^{n} h^{\tau}_{\gamma}(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}) - 2\rho_{\tau}(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}) = o(1). \tag{2}$$

By the Basic Corollary of Lemma 2 from Hjort and Pollard (1993), $\hat{\boldsymbol{\beta}}^{\tau}_{\gamma}$ has the same limiting distribution as the standard quantile regression estimator which completes the proof. □

Proof of Lemma 1

Proof. Note that $h^{\tau}_{\gamma}(u) - 2\rho_{\tau}(u) = h_{\gamma}(u) - |u|$. Also, $|u| - \gamma/2 \leq h_{\gamma}(u) \leq |u|$ which implies $h_{\gamma}(u) - |u| \geq -\gamma/2$ and $|u| - h_{\gamma}(u) \geq 0$. Combining both provides $h^{\tau}_{\gamma}(u) - 2\rho_{\tau}(u) \geq -\gamma/2$ and $2\rho_{\tau}(u) - h^{\tau}_{\gamma}(u) \geq 0$. Thus under the assumption $H^{\tau}_{\gamma,\lambda}(\boldsymbol{\beta}) \leq H^{\tau}_{\gamma,\lambda}(\boldsymbol{\beta}^{*}_{\tau})$,

$$0 \geq \frac{1}{n}\sum_{i=1}^{n}\left\{\rho_{\tau}(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}) - \rho_{\tau}[y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}^{*}_{\tau}]\right\} + \lambda\sum_{g=1}^{G}\sqrt{d_g}\left(||\boldsymbol{\beta}_g||_2 - ||\boldsymbol{\beta}^{*}_{\tau g}||_2\right) - \gamma/2.$$

By convexity, and definition of subgradients,

$$\rho_{\tau}(y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}) - \rho_{\tau}[y_i - \mathbf{x}_i^{\top}\boldsymbol{\beta}^{*}_{\tau}] \geq -\left\{\tau - I[y_i \leq \mathbf{x}_i^{\top}\boldsymbol{\beta}^{*}_{\tau}]\right\}\mathbf{x}_i^{\top}[\boldsymbol{\beta} - \boldsymbol{\beta}^{*}_{\tau}].$$

Combining this with the previous inequality and the assumption that $\lambda \geq 2\Lambda_\tau$,

$$\begin{aligned} 0 \quad \geq \quad & -\Lambda_\tau \sum_{g=0}^{q} \sqrt{d_g} \left(||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2 \right) + \lambda \sum_{g=1}^{q} \sqrt{d_g} \left(||\boldsymbol{\beta}_g||_2 - ||\boldsymbol{\beta}^*_{\tau g}||_2 \right) \\ & -\Lambda_\tau \sum_{g=q+1}^{G} \sqrt{d_g} ||\boldsymbol{\beta}_g||_2 + \lambda \sum_{g=q+1}^{G} \sqrt{d_g} ||\boldsymbol{\beta}_g||_2 - \gamma/2. \end{aligned}$$

By the triangle inequality $-||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2 \leq ||\boldsymbol{\beta}_g||_2 - ||\boldsymbol{\beta}^*_{\tau g}||_2$. Additionally, $-\lambda \sum_{g=1}^{q} \sqrt{d_g} ||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2 \geq -\lambda \sum_{g=0}^{q} \sqrt{d_g} ||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2$. Thus,

$$0 \quad \geq \quad -(\Lambda_\tau + \lambda) \sum_{g=0}^{q} \sqrt{d_g} \left(||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2 \right) + (\lambda - \Lambda_\tau) \sum_{g=q+1}^{G} \sqrt{d_g} ||\boldsymbol{\beta}_g||_2 - \gamma/2.$$

Therefore,

$$(\lambda - \Lambda_\tau) \sum_{g=q+1}^{G} \sqrt{d_g} ||\boldsymbol{\beta}_g||_2 \leq (\Lambda_\tau + \lambda) \sum_{g=0}^{q} \sqrt{d_g} \left(||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2 \right) + \gamma/2. \tag{3}$$

If $\lambda \geq 2\Lambda_\tau$ then $\Lambda_\tau + \lambda \leq \lambda/2 + \lambda = 3/2\lambda$ and $\lambda - \Lambda_\tau \geq \lambda - \lambda/2 = \lambda/2$. Therefore, $\lambda/2 \sum_{g=q+1}^{G} \sqrt{d_g} ||\boldsymbol{\beta}_g||_2 \leq 3/2\lambda \sum_{g=0}^{q} \sqrt{d_g} \left(||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2 \right) + \gamma/2$ and $\sum_{g=q+1}^{G} \sqrt{d_g} ||\boldsymbol{\beta}_g||_2 \leq 3 \sum_{g=0}^{q} \sqrt{d_g} \left(||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2 \right) + \gamma/\lambda$. Thus completing the proof. □

Results for proof of Theorem 2

The following is Corollary A.1 from Kato (2011) and repeated here for ease of presentation. Other results are similar to those from Kato (2011), but different due to the use of the Huber approximation to the quantile loss.

Corollary 1. *Let $b_1, \ldots, b_n$ be independent Rademacher random variables and independent of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Define $Z = || \sum_{i=1}^{n} b_i \mathbf{x}_i ||_2$, and $\zeta = \sup_{||\mathbf{a}||=1} \mathbf{a}^\top \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{a}$ then for any positive $\eta > 0$*

$$E[exp(\eta Z)] \leq 16 exp \left[\eta \sqrt{2E[Z^2]} + 4\eta^2 \zeta \right].$$

Define $\mathcal{B}_\delta = \left\{ \boldsymbol{\beta} \in \mathbb{R}^p | \boldsymbol{\beta} - \boldsymbol{\beta}^* \in \mathcal{C}^\tau_{\gamma,\lambda}, ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2 = \delta \right\}$. Let $\mathbf{z}_i = (y_i, \mathbf{z}_i^\top)^\top$ and define $m^\tau_{\gamma,\boldsymbol{\beta}}(\mathbf{z}_i) = h^\tau_\gamma(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - h^\tau_\gamma(\epsilon^\tau_i)$ and $m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i) = \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) - \rho_\tau(\epsilon^\tau_i)$. The next three lemmas are used in the proof of Theorem 2.

Lemma 1. *For any $\delta \in (\gamma/\lambda, C_r)$ if $||\hat{\boldsymbol{\beta}}^\tau_{\gamma,\lambda} - \boldsymbol{\beta}^*_\tau||_2 \geq \delta$, $\lambda \geq 2\Lambda_\tau$ and Conditions 1 - 5 hold then*

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}_\delta} \left| \sum_{i=1}^{n} \left\{ m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i) - E\left[m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i) \right] \right\} \right| > \delta(c_f/3\phi^2_{\min} n\delta - \sqrt{d^*} n\lambda - n\lambda/2). \tag{4}$$

Proof. Lemma 1 provides that if $\lambda \geq 2\Lambda_\tau$ then $\hat{\boldsymbol{\beta}}_{\gamma,\lambda} - \boldsymbol{\beta}^* \in \mathcal{B}_\delta$. Therefore under the assumptions of this lemma there exists $\boldsymbol{\beta} \in \mathcal{B}_\delta$ such that

$$\begin{aligned} 0 &\geq \frac{1}{2}\sum_{i=1}^{n} m^{\tau}_{\gamma,\boldsymbol{\beta}}(\mathbf{z}_i) + n\lambda \sum_{g=1}^{G} \sqrt{d_g}\left(||\boldsymbol{\beta}_g||_2 - ||\boldsymbol{\beta}^*_{\tau g}||_2\right) \\ &\geq \sum_{i=1}^{n} m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z}_i) - n\lambda \sum_{g=1}^{q} \sqrt{d_g}||\boldsymbol{\beta}_g - \boldsymbol{\beta}^*_{\tau g}||_2 - n\delta\lambda/2 \\ &\geq \sum_{i=1}^{n} m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z}_i) - n\sqrt{d^*}\lambda\delta - n\delta\lambda/2. \end{aligned}$$

Note, $\sum_{i=1}^{n} m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z}_i) \geq nE\left[m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z})\right] - \sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^{n} m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z}_i) - E\left[m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z}_i)\right]\right|$ and thus proof can be completed by deriving satisfactory lower bounds for $nE\left[m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z})\right]$. Knight's identity, for $\tau = 1/2$ see (Knight, 1998) and generalized for any τ in (Koenker, 2005), provides that

$$\rho_\tau(u-v) - \rho_\tau(u) = -v\left[\tau - I(u \leq 0)\right] + \int_0^v \left[I(u \leq s) - I(u \leq 0)\right] ds.$$

As defined in Condition 5,

$$C_r = \frac{c_f}{C'_f \phi_{\max}} \inf_{\boldsymbol{\alpha}\in\mathcal{C}^{\tau}_{\gamma,\lambda}, ||\boldsymbol{\alpha}||_2=1} \frac{E[(\boldsymbol{\alpha}^\top \mathbf{x}_i)^2]^{3/2}}{E[|\boldsymbol{\alpha}^\top \mathbf{x}|^3]},$$

and therefore

$$\frac{c_f}{C_r C'_f \phi_{\max}} = \sup_{\boldsymbol{\alpha}\in\mathcal{C}^{\tau}_{\gamma,\lambda}, ||\boldsymbol{\alpha}||_2=1} \frac{E[|\boldsymbol{\alpha}^\top \mathbf{x}|^3]}{E[(\boldsymbol{\alpha}^\top \mathbf{x}_i)^2]^{3/2}}. \tag{5}$$

For $\delta \in (\gamma/\lambda, C_r)$ it follows that

$$\begin{aligned} E\left[m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z})\right] &\geq c_f/2E\left(\left\{\mathbf{x}^\top[\boldsymbol{\beta} - \boldsymbol{\beta}^*_\tau]\right\}^2\right) - 1/6C'_f E\left(\left|\left\{\mathbf{x}^\top[\boldsymbol{\beta} - \boldsymbol{\beta}^*_\tau]\right\}\right|^3\right) \\ &\geq c_f/2E\left(\left\{\mathbf{x}_1^\top[\boldsymbol{\beta} - \boldsymbol{\beta}^*_\tau]\right\}^2\right) - \frac{c_f\delta}{6C_r}E\left(\left\{\mathbf{x}^\top[\boldsymbol{\beta} - \boldsymbol{\beta}^*_\tau]\right\}^2\right) \geq c_f/3\phi^2_{\min}\delta^2. \end{aligned}$$

Which, completes the proof.

□

Lemma 2. *Assuming Conditions 1-5 hold, for any $\delta > \gamma/\lambda$ and $t \geq \phi_{\max}\delta\sqrt{8n}$*

$$\begin{aligned} & P\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^n m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i) - E\left[m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i)\right]\right| \geq t\right) \\ \leq \;& 64G\,exp\left\{-\frac{[t/\delta - 8(4\sqrt{d^*+1}+1)\sqrt{3n}]^2}{2560(4\sqrt{d^*+1}+1)^2 n/d_{\min}}\right\} + 4\kappa. \end{aligned}$$

Proof. Define $M_{\boldsymbol{\beta}}(\mathbf{z}_i) = \sum_{i=1}^n m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i) - E\left[m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i)\right]$. Let $b_1,\ldots,b_n$ be independent Rademacher random variables that are also independent of $\mathbf{z}_1,\ldots,\mathbf{z}_n$. Define $P_{\mathbf{b}}$ and $E_{\mathbf{b}}$ as the conditional probability and conditional expectation of $b_1,\ldots,b_n$ given $\mathbf{z}_1,\ldots,\mathbf{z}_n$. By Condition 3 it follows that for $\boldsymbol{\beta}\in\mathcal{B}_\delta$ that $E[m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i)^2] \leq \phi^2_{\max}\delta^2$. Thus by Lemma 2.3.7 from van der Vaart and Wellner (1996) and the assumption that $t \geq \phi_{\max}\delta\sqrt{8n}$ it follows that

$$\begin{aligned} P\left(\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}|M_{\boldsymbol{\beta}}(\mathbf{z}_i)| > t\right) &\leq 4E\left\{P_{\mathbf{b}}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^n b_i m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i)\right| > t/4\right]\right\} \\ &\leq 4E\left\{P_{\mathbf{b}}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^n b_i m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i)\right| > t/4\right]\middle|\Omega_0\right\} + 4\kappa. \end{aligned}$$

For any $s > 0$, by Markov's inequality

$$P_{\mathbf{b}}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^n b_i m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i)\right| > t/4\middle|\Omega_0\right] \leq \exp(-st/4)E_{\mathbf{b}}\left\{\exp\left[s\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^n b_i m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i)\right|\right]\middle|\Omega_0\right\}.$$

Define $w_{\boldsymbol{\beta}}(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\beta}$, by Theorem 4.12 from Ledoux and Talagrand (1991)

$$E_{\mathbf{b}}\left\{\exp\left[s\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^n b_i m^\tau_{\boldsymbol{\beta}}(\mathbf{z}_i)\right|\right]\middle|\Omega_0\right\} \leq E_{\mathbf{b}}\left\{\exp\left[2s\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^n b_i w_{\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau}(\mathbf{x}_i)\right|\right]\middle|\Omega_0\right\}.$$

Note, $\left|\sum_{i=1}^n b_i w_{\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau}(\mathbf{x}_i)\right| = \left|\sum_{g=0}^G \sqrt{d_g}(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g^\top\left(\sum_{i=1}^n b_i\mathbf{x}_{ig}/\sqrt{d_g}\right)\right|$. Define $Z_g = ||\sum_{i=1}^n b_i\mathbf{x}_{ig}/\sqrt{d_g}||_2$. By the triangle inequality and using max to pull Z_g out of the sum we have

$$\begin{aligned} & \sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{g=0}^G \sqrt{d_g}(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g^\top\left(\sum_{i=1}^n b_i\mathbf{x}_{ig}/\sqrt{d_g}\right)\right| \\ \leq \;& \left(\max_{g\in\{0,\ldots,G\}} Z_g\right)\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left(\sum_{g=0}^q \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2 + \sum_{g=q+1}^G \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2\right). \end{aligned}$$

By definition of $\mathcal{B}_\delta$

$$\begin{aligned}
&\left(\max_{g\in\{0,\ldots,G\}} Z_g\right) \sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta} \left(\sum_{g=0}^{q} \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2 + \sum_{g=q+1}^{G} \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2\right) \\
\leq\ &\left(\max_{g\in\{0,\ldots,G\}} Z_g\right) \sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta} \left[\sum_{g=0}^{q} \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2 + \left(3\sum_{g=0}^{q} \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2 + \gamma/\lambda\right)\right].
\end{aligned}$$

By the Cauchy-Schwarz inequality

$$\sum_{g=0}^{q} \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2 \leq \sqrt{\sum_{g=0}^{q} d_g}\sqrt{\sum_{g=0}^{q} ||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2^2} \leq \sqrt{d^*+1}\delta.$$

Using the above inequality, the definition of $\mathcal{B}_\delta$ and that $\delta > \gamma/\lambda$ we have

$$\begin{aligned}
&\left(\max_{g\in\{0,\ldots,G\}} Z_g\right) \sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta} \left[\sum_{g=0}^{q} \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2 + \left(3\sum_{g=0}^{q} \sqrt{d_g}||(\boldsymbol{\beta}-\boldsymbol{\beta}^*_\tau)_g||_2 + \gamma/\lambda\right)\right] \\
\leq\ &\delta(4\sqrt{d^*+1}+1)\max_{0\leq g\leq G} Z_g.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&E_{\mathbf{b}}\left\{\exp\left[2s\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^{n} b_i w_{\boldsymbol{\beta}_-\boldsymbol{\beta}^*_\tau}(\mathbf{x}_i)\right|\right]\middle|\Omega_0\right\} \\
\leq\ &E_{\mathbf{b}}\left\{\exp\left[2s\delta(4\sqrt{d^*+1}+1)\max_{0\leq g\leq G} Z_g\right]\middle|\Omega_0\right\} \\
\leq\ &\sum_{g=0}^{G} E_{\mathbf{b}}\left\{\exp\left[2s\delta(4\sqrt{d^*+1}+1)Z_g\right]\middle|\Omega_0\right\}.
\end{aligned}$$

Under the event Ω_0 and noting that $\mathbf{x}_i^\top\mathbf{x}_i = \mathrm{Tr}(\mathbf{x}_i^\top\mathbf{x}_i) = \mathrm{Tr}(\mathbf{x}_i\mathbf{x}_i^\top)$ it follows that for any $g \in \{0,\ldots,G\}$

$$\sup_{||\mathbf{a}||_2=1} d_g^{-1}\mathbf{a}^\top\sum_{i=1}^{n}\mathbf{x}_{ig}\mathbf{x}_{ig}^\top\mathbf{a} = d_g^{-1}n||\hat{\Sigma}_g||^2 \leq 2d_g^{-1}n(||I_{d_g}||^2 + ||\hat{\Sigma}_g - I_{d_g}||^2) \leq d_g^{-1}n2.5,$$

and

$$E_{\mathbf{b}}[Z_g^2|\Omega_0] = d_g^{-1}\sum_{i=1}^{n}\mathbf{x}_{ig}^\top\mathbf{x}_{ig} = d_g^{-1}n\mathrm{Tr}(\hat{\Sigma}_g) \leq d_g^{-1}nd_g\lambda_{\max}(\hat{\Sigma}_g) \leq 1.5n.$$

Therefore by Corollary 1, for any positive value C

$$E_{\mathbf{b}}[\exp(CZ_g)|\Omega_0] \leq 16\exp\left(C\sqrt{3n} + 10C^2n/d_g\right),$$

and

$$\begin{aligned}
&\sum_{g=0}^{G} E_{\mathbf{b}}\left\{\exp\left[2s\delta(4\sqrt{d^*+1}+1)Z_g\right]\Big|\Omega_0\right\} \\
\leq\ & 16G\exp\left[2s\delta(4\sqrt{d^*+1}+1)\sqrt{3n}+40s^2\delta^2(4\sqrt{d^*+1}+1)^2 n/d_{\min}\right].
\end{aligned}$$

Therefore, for $A = 2\delta(4\sqrt{d^*+1}+1)\sqrt{3n}$ and $B = 40\delta^2(4\sqrt{d^*+1}+1)^2 n/d_{\min}$,

$$\begin{aligned}
&P_{\mathbf{b}}\left[\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^{n} b_i m^{\tau}_{\boldsymbol{\beta}}(\mathbf{z}_i)\right| > t/4\Big|\Omega_0\right] \\
\leq\ & 16G\exp\left(As+Bs^2\right)\exp(-st/4) = 16G\exp\left[s(A-t/4)+Bs^2\right].
\end{aligned}$$

Function is minimized at $s = \frac{t-4A}{8B}$, and thus we have the following upper bound for the exponential function

$$\exp\left[\frac{t-4A}{8B}(A-t/4)+\frac{(t-4A)^2}{64B}\right] = \exp\left\{-\frac{[t/\delta-8(4\sqrt{d^*+1}+1)\sqrt{3n}]^2}{2560(4\sqrt{d^*+1}+1)^2 n/d_{\min}}\right\}.$$

This completes the proof. □

The following lemma is very similar to Lemma A.4 from Kato (2011), but has been slightly modified to better match the notation and setting of this paper.

Lemma 3. *For $t_1 > 0$ and $t_2 > 0$ then*

$$P\left(\Lambda_\tau > 4\sqrt{2/n}+t_1/n+t_2/n|\mathbf{z}_1^n\right) \leq 2exp\left[-t_1^2/(2n)\right]+16Gexp\left[-d_{\min}t_2^2/(128n)\right].$$

Proof. Proof follows from proof of Lemma A.4 of Kato (2011) and noticing that $n^{-1}\Lambda$ from that paper is equal to Λ_τ in this paper and our G is equal to their $q-1$. □

Proof of Theorem 2

Proof. Set

$$\begin{aligned}
\delta &= \frac{3}{c_f\phi^2_{\min}}\left[\max\left(\delta^*,\sqrt{\frac{8\phi^2_{\max}}{n}}\right)+\lambda(\sqrt{d^*}+1/2)\right], \\
t &= \delta\left[c_f/3\phi^2_{\min}n\delta-n\lambda(\sqrt{d^*}+1/2)\right].
\end{aligned}$$

By construction $\delta < \gamma/\lambda$ and by assumption, for sufficiently large n $\delta < C_r$. Thus Lemma 1 holds and

$$\begin{aligned} & P\left(||\hat{\boldsymbol{\beta}}_{\gamma}^{\tau} - \boldsymbol{\beta}_{\tau}^{*}||_2 \geq \delta, \lambda \geq 2\Lambda_\tau\right) \\ \leq \; & P\left\{\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^{n} m_{\boldsymbol{\beta}}^{\tau}(\mathbf{z}_i) - E\left[m_{\boldsymbol{\beta}}^{\tau}(\mathbf{z}_i)\right]\right| > \delta(c_f/3\phi_{\min}^2 n\delta - \sqrt{d^*}n\lambda - n\lambda/2)\right\}. \end{aligned}$$

In addition, $t \geq \phi_{\max}\delta\sqrt{8n}$ and thus Lemma 2 holds. Using the upper bound provided by Lemma 2 and the stated values of δ and t

$$P\left\{\sup_{\boldsymbol{\beta}\in\mathcal{B}_\delta}\left|\sum_{i=1}^{n} m_{\boldsymbol{\beta}}^{\tau}(\mathbf{z}_i) - E\left[m_{\boldsymbol{\beta}}^{\tau}(\mathbf{z}_i)\right]\right| > t\right\} \leq 64G^{1-C_1^2} + 4\kappa.$$

Therefore,

$$P\left(||\hat{\boldsymbol{\beta}}_{\tau} - \boldsymbol{\beta}_{\tau}^{*}||_2 \geq \delta\right) \leq P(\lambda < 2\Lambda_\tau) + 64G^{1-C_1^2} + 4\kappa.$$

To complete the proof, by Lemma 3 with $t_1 = \sqrt{n}C_2$ and $t_2 = C_3\sqrt{\frac{n\log(G)}{d_{\min}}}$

$$\begin{aligned} P(\Lambda_\tau > \lambda/2) &= P\left(\Lambda_\tau > \frac{(4\sqrt{2}+C_2)}{\sqrt{n}} + C_3\sqrt{\frac{\log(G)}{nd_{\min}}}\right) \\ &\leq 2\exp(-C_2^2/2) + 16G^{1-C_3^2/128}. \end{aligned}$$

Thus completing the proof of Theorem 2. □

Proof of Corollary 1

Proof. Using Theorem 2 with $C_1 = t$ we have with probability going to one that

$$||\hat{\boldsymbol{\beta}}_{\gamma,\lambda}^{\tau} - \boldsymbol{\beta}_{\tau}^{*}||_2 \leq \frac{3}{c_f\phi_{\min}^2}\left[\max\left(\delta^*, \sqrt{\frac{8\phi_{\max}^2}{n}}\right) + \lambda(\sqrt{d^*} + 1/2)\right].$$

By Condition 3 the dominating term is $\lambda\sqrt{d^*}$. Proof is complete because under the stated conditions $\lambda\sqrt{d^*} = O\left\{t\sqrt{\frac{d^*}{n}\left[1 + \frac{\log(G)}{d_{\min}}\right]}\right\}$. □

Proof of Corollary 2

Proof. Using Theorem 2 with $C_1 = \sqrt{2}$, we have with probability at least $1 - 2\exp(-t^2/2) + 16G^{1-T^2/128} + 64G^{-1} + 4\kappa$ that

$$||\hat{\boldsymbol{\beta}}_{\gamma,\lambda}^{\tau} - \boldsymbol{\beta}_{\tau}^{*}||_2 \leq \frac{3}{c_f\phi_{\min}^2}\left[\max\left(\delta^*, \sqrt{\frac{8\phi_{\max}^2}{n}}\right) + \lambda(\sqrt{d^*} + 1/2)\right].$$

By Condition 3 the dominating term is $\lambda\sqrt{d^*}$ and under the conditions of the corollary

$$2\exp(-t^2/2) + 16G^{1-T^2/128} + 64G^{-1} + 4\kappa \to 0.$$

Proof is complete by noticing $\lambda\sqrt{d^*} = O\left\{\sqrt{\frac{d^*}{n}\left[t + \frac{\log(G)}{d_{\min}}\right]}\right\}$.

□

Derivation of equation (8)

Let

$$\hat{\boldsymbol{\beta}}_g^* = \arg\min_{\boldsymbol{\beta}_g} Q_{\gamma,\tau}(\boldsymbol{\beta}_g; \tilde{\boldsymbol{\beta}}_{-g}) + \lambda\sqrt{d_g}\|\boldsymbol{\beta}_g\|, \tag{6}$$

where

$$Q_{\gamma,\tau}(\boldsymbol{\beta}_g; \tilde{\boldsymbol{\beta}}_{-g}) = L_{\gamma,\tau}(\tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta}_g - \tilde{\boldsymbol{\beta}}_g)^\top \nabla_g L_{\gamma,\tau}(\tilde{\boldsymbol{\beta}}) + \frac{\xi_g}{2}(\boldsymbol{\beta}_g - \tilde{\boldsymbol{\beta}}_g)^\top(\boldsymbol{\beta}_g - \tilde{\boldsymbol{\beta}}_g).$$

Denote $\boldsymbol{v}_g = \nabla_g L_{\gamma,\tau}(\tilde{\boldsymbol{\beta}})$. For $\|\boldsymbol{\beta}_g\| \neq 0$, the KKT condition of (6) is

$$\boldsymbol{v}_g + \xi_g(\boldsymbol{\beta}_g - \tilde{\boldsymbol{\beta}}_g) + \frac{\lambda\sqrt{d_g}}{\|\boldsymbol{\beta}_g\|}\boldsymbol{\beta}_g = \mathbf{0}.$$

Therefore,

$$\left(\frac{\lambda\sqrt{d_g}}{\|\boldsymbol{\beta}_g\|} + \xi_g\right)\boldsymbol{\beta}_g = \xi_g\tilde{\boldsymbol{\beta}}_g - v_g; \tag{7}$$

$$\left(\frac{\lambda\sqrt{d_g}}{\|\boldsymbol{\beta}_g\|} + \xi_g\right)\|\boldsymbol{\beta}_g\| = \|\xi_g\tilde{\boldsymbol{\beta}}_g - v_g\|; \tag{8}$$

$$\|\boldsymbol{\beta}_g\| = \frac{1}{\xi_g}\left(\|\xi_g\tilde{\boldsymbol{\beta}}_g - v_g\| - \lambda\sqrt{d_g}\right). \tag{9}$$

Substituting (9) to (7), we obtain

$$\boldsymbol{\beta}_g = \frac{1}{\xi_g}\left(\xi_g\tilde{\boldsymbol{\beta}}_g - v_g\right)\left(1 - \frac{\lambda\sqrt{d_g}}{\|\xi_g\tilde{\boldsymbol{\beta}}_g - v_g\|}\right).$$

Note that $\|\xi_g\tilde{\boldsymbol{\beta}}_g - v_g\| > \lambda\sqrt{d_g}$ holds because (8) implies $\mathrm{sgn}(\boldsymbol{\beta}_g) = \mathrm{sgn}(\xi_g\tilde{\boldsymbol{\beta}}_g - v_g)$. Hence,

$$\boldsymbol{\beta}_g = \frac{1}{\xi_g}\left(\xi_g\tilde{\boldsymbol{\beta}}_g - v_g\right)\left(1 - \frac{\lambda\sqrt{d_g}}{\|\xi_g\tilde{\boldsymbol{\beta}}_g - v_g\|}\right)_+.$$

For $\|\boldsymbol{\beta}_g\| = 0$, the KKT condition is

$$v_g + \xi_g(\boldsymbol{\beta}_g - \tilde{\boldsymbol{\beta}}_g) + \lambda\sqrt{d_g}s = 0,$$

where $\|s\| \leq 1$. This implies that $\|\xi_g\tilde{\boldsymbol{\beta}}_g - v_g\| \leq \lambda\sqrt{d_g}$.

References

Hjort, N. L. and D. Pollard (1993). Asymptotics for minimisers of convex processes. Technical report, University of Oslo and Yale University.

Kato, K. (2011). Group lasso for high dimensional sparse quantile regression models.

Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *Ann. Statist. 26*(2), 755–770.

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

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces*. Springer-Verlag.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Cambridge University Press.