Statistica Sinica **33** (2023), 945-959
doi:https://doi.org/10.5705/ss.202021.0140

# SPARSE AND LOW-RANK MATRIX QUANTILE ESTIMATION WITH APPLICATION TO QUADRATIC REGRESSION

Wenqi Lu[1,2,3], Zhongyi Zhu[2] and Heng Lian[3,4]

[1]*Nankai University,* [2]*Fudan University,*
[3]*City University of Hong Kong and* [4]*CityU Shenzhen Research Institute*

*Abstract:* This study examines matrix quantile regression where the covariate is a matrix and the response is a scalar. Although the statistical estimation of matrix regression is an active field of research, few studies examine quantile regression with matrix covariates. We propose an estimation procedure based on convex regularizations in a high-dimensional setting. In order to reduce the dimensionality, the coefficient matrix is assumed to be low rank and/or sparse. Thus, we impose two regularizers to encourage different low-dimensional structures. We develop the asymptotic properties and an implementation based on the incremental proximal gradient algorithm. We then apply the proposed estimator to quadratic quantile regression, and demonstrate its advantages using simulations and a real-data analysis.

*Key words and phrases:* Dual norm, interaction effects, matrix regression, penalization.

## 1. Introduction

Quantile regression (Koenker and Bassett (1978)) is a useful statistical tool in data analysis. It provides a complement to a mean regression, allowing us to analyze the entire conditional distribution by modeling the covariate effects at different quantile levels. Despite there being a large body of literature on the theoretical and computational aspects of vector covariate quantile regression (Koenker (2005); Belloni and Chernozhukov (2011); Yu, Lin and Wang (2017); Yi and Huang (2017)), matrix quantile regression is rarely studied. However, matrix data arise frequently in fields such as digital image analysis (Zhou and Li (2014)), multi-task regression (Yuan et al. (2007); Argyriou, Evgeniou and Pontil (2008); Bunea, She and Wegkamp (2012)), matrix completion (Candes and Plan (2010); Koltchinskii, Lounici and Tsybakov (2011); Negahban and Wainwright (2012)), and quadratic regression (Bien, Taylor and Tibshirani (2013)).

Corresponding author: Heng Lian, Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China. E-mail: henglian@cityu.edu.hk.

The primary challenge in matrix data analysis is its typically high-dimensional nature. A popular way to reduce the dimensionality is to impose a sparsity assumption on the covariates, which is often encouraged by penalties such as the lasso (Tibshirani (1996)), smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)), elastic net (Zou and Hastie (2005)), and many others. For high-dimensional vector quantile regression with sparsity assumptions, Belloni and Chernozhukov (2011) established a uniform convergence rate for $\ell_1$ penalization. Later, Zheng, Peng and He (2015) achieved the oracle rate by employing an adaptive lasso penalty. Other recent works studying related problems include, among others, Kato (2011), Chao, Volgushev and Cheng (2017), Belloni et al. (2019), and Pan and Zhou (2020).

For matrix data, a low-dimensional structure can be in the form of sparsity and/or low rankness. The nuclear norm is a convex relaxation of the matrix rank, so it is used as a penalty in many penalized least squares approaches to encourage low rankness (Yuan et al. (2007); Argyriou, Micchelli and Pontil (2010); Koltchinskii, Lounici and Tsybakov (2011); Negahban and Wainwright (2011); Zhou and Li (2014)). Other penalties, such as the rank (Bunea, She and Wegkamp (2011)), Von Neumann entropy (Koltchinskii (2011)), and Schatten-p norm (Rohde and Tsybakov (2011)) are also used. Furthermore, some works consider low rankness and sparsity to further improve the dimension reduction or interpretation. For example, Agarwal, Negahban and Wainwright (2012) decomposed the true signal into a sum of a low-rank matrix and a sparse matrix. Other works assume a coefficient matrix satisfying low rankness and sparsity simultaneously, such as the sparse reduced-rank regression (Chen, Chan and Stenseth (2012); Ma, Ma and Sun (2020)) and two-step joint rank and row selection estimator (Bunea, She and Wegkamp (2012)). However, these works are all based on penalized least squares.

We propose an estimator in quantile regression with matrix covariates and a scalar response in a high-dimensional setting. Compared with mean regression, quantile regression has advantages in terms of its robustness to outliers, skewness, and heterogeneity, and it can be used to build prediction intervals. In order to deal with the high dimensionality, we apply convex regularization techniques. In particular, we assume the underlying matrix lies in a low-dimensional subspace that is both sparse and low rank. Then, we provide a convex regularized optimization approach using both the nuclear norm and the entry-wise $\ell_1$ norm as regularizers to exploit the low-dimensional structure. Unlike some previous approaches, our method encourages low rankness and sparsity simultaneously. Moreover, we derive the upper bound on the estimation error of the proposed method in the high-dimensional setting. Theoretical results for high-dimensional quantile regression are more complicated than those of the least squares regression models. They also require more technical analysis associated with the matrix norms than in the case of penalized quantile regression with vector coefficients.

We then apply the matrix quantile regression to linear quantile regression with interaction effects. Dimension reduction is desirable for models with interactions, because even when the number of covariates $p$ is moderate, quadratic regression involves $O(p^2)$ parameters. Several variable selection methods have been proposed to reduce the number of parameters for quadratic regression, including regularization methods (Choi, Li and Zhu (2010); Bien, Taylor and Tibshirani (2013); Hao, Feng and Zhang (2018)) and screening (Hao and Zhang (2014); Fan et al. (2015)). These works all rely on the sparsity assumption, which requires that the number of significant variables is small and the signal size is sufficiently large. We consider an alternative strategy using matrix regression, which does not necessarily require sparsity. Note that by writing $\mathbf{Z}_i = (1, \mathbf{x}_i^\top)^\top(1, \mathbf{x}_i^\top)$, where $\mathbf{x}_i$ is a $p$-dimensional vector predictor, the main effect $\mathbf{x}_i$ and quadratic interactions are all incorporated in matrix form. Thus, a rank constraint can be used to restrict the effective number of parameters.

The rest of the paper is organized as follows. In Section 2, we introduce the estimator for the matrix quantile regression model based on regularization, and present the implementation details and application to quadratic regression. Section 3 establishes the theoretical properties. In Section 4, we investigate the finite-sample properties on simulated and real data sets in quadratic quantile regression. We conclude the paper in Section 5.

## 2. Matrix Quantile Regression

### 2.1. General model setup

In this paper, we study a matrix quantile regression model with a scalar response $y \in \mathbb{R}$ and a matrix covariate $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$. Define the $\tau$th conditional quantile of $y$ given $\mathbf{Z}$ as $Q_\tau(y|\mathbf{Z}) = \inf\{t : F_{y|\mathbf{Z}}(t) \geq \tau\}$, where $F_{y|\mathbf{Z}}(t)$ is the conditional distribution function. We consider the setting that, for a certain quantile level $\tau \in (0,1)$, $Q_\tau(y|\mathbf{Z})$ is modeled by the linear regression model

$$Q_\tau(y|\mathbf{Z}) = \langle \mathbf{B}, \mathbf{Z} \rangle, \tag{2.1}$$

where $\mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$ and $\langle \mathbf{B}, \mathbf{Z} \rangle = \mathrm{tr}(\mathbf{B}^\top \mathbf{Z}) = \langle \mathrm{vec}(\mathbf{B}), \mathrm{vec}(\mathbf{Z}) \rangle$ is the inner product between matrices. In the above, we omit the intercept for simplicity. The intercept does not play a significant role in developing the theory, but is certainly useful in practice. On the other hand, the intercept is already incorporated into $\mathbf{B}$ for quadratic regression, and thus in such a special case, an additional intercept in (2.1) is not necessary.

We apply the convex regularization framework to estimate the coefficient $\mathbf{B}$ under low-dimensionality assumptions, including low rankness and sparsity. Given an independent and identically distributed (i.i.d.) sample $(y_i, \mathbf{Z}_i)$, for $i = 1, \ldots, n$, the regularized estimator is defined by

$$\widehat{\mathbf{B}} = \operatorname{argmin} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \langle \mathbf{B}, \mathbf{Z}_i \rangle) + \lambda_1 \mathcal{R}_1(\mathbf{B}) + \lambda_2 \mathcal{R}_2(\mathbf{B}), \tag{2.2}$$

where $\rho_\tau(u) = u(\tau - I\{u < 0\})$ is the check loss function, and $\mathcal{R}_1(\mathbf{B})$ and $\mathcal{R}_2(\mathbf{B})$ are the regularizers that exploit the low rankness and sparsity structure, respectively. Let $(\sigma_1(\mathbf{B}), \ldots, \sigma_r(\mathbf{B}))$ be the nonzero singular values of $\mathbf{B}$, with $r = \operatorname{rank}(\mathbf{B})$ the rank of $\mathbf{B}$. The nuclear norm $\|\mathbf{B}\|_* = \sum_{j=1}^{r} \sigma_j(\mathbf{B})$ is a convex relaxation of $\operatorname{rank}(\mathbf{B})$. Thus, we use $\mathcal{R}_1(\mathbf{B}) = \|\mathbf{B}\|_*$ to encourage low rankness. A widely used regularizer to encourage entry-wise sparsity is the $\ell_1$ norm, such as the lasso in classical linear regression (Tibshirani (1996)). We use $\mathcal{R}_2(\mathbf{B}) = \|\mathbf{B}\|_1 := \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} |\mathbf{B}_{jk}|$ as the sparsity regularizer.

The convex optimization problem (2.2) includes two regularizers, and the optimization problem with one penalty can be solved using a proximal gradient algorithm. Thus, we can use the *incremental* proximal gradient method (Bertsekas (2011)). Specifically, denoting $\ell(\mathbf{B}) = (1/n) \sum_{i=1}^{n} \rho_\tau(y_i - \langle \mathbf{B}, \mathbf{Z}_i \rangle)$, the incremental proximal gradient method operates on $\mathcal{R}_1$ and $\mathcal{R}_2$ in turn, and treats $\ell(\mathbf{B})$ in a (sub-)gradient step. The $t$th iteration of the algorithm computes

$$\begin{aligned}
\mathbf{B}_1^t &= \operatorname{argmin} \left\{ \mathcal{R}_1(\mathbf{B}) + \frac{1}{2\gamma} \|\mathbf{B} - \mathbf{B}^{t-1}\|_F^2 \right\}, \\
\mathbf{B}_2^t &= \operatorname{argmin} \left\{ \mathcal{R}_2(\mathbf{B}) + \frac{1}{2\gamma} \|\mathbf{B} - \mathbf{B}_1^{t}\|_F^2 \right\}, \\
\mathbf{B}^t &= \mathbf{B}_2^t - \gamma \nabla \ell(\mathbf{B}_2^t),
\end{aligned}$$

where $\nabla \ell(\mathbf{B})$ is a sub-derivative of the loss, and $\gamma$ is the step size. The pseudocode is presented in Algorithm 1. The initial value $\mathbf{B}^0$ is a matrix with independent standard normal entries. In fact, because the optimization problem is convex, the initial estimator has little effect in our procedure. For the step size, setting $\gamma$ too large may make the algorithm fail to converge, while too small a value makes the convergence very slow. In our simulations, the step size $\gamma$ is set to 0.1, which is satisfactory in our numerical studies. An investigation of a more principled and adaptive approach for the step size is left for future work. We stop the algorithm when the decrease of the objective function value is less than $10^{-5}$. Because the algorithm can be seen as a special case of the incremental proximal gradient method, its numerical convergence is guaranteed by Proposition 3 and Proposition 4 in Bertsekas (2011).

**Algorithm 1.** Incremental proximal gradient method for quantile matrix regression.

**Input:** Initial value $\mathbf{B}^0$, $\gamma$
**repeat**
SVD for $\mathbf{B}^{t-1}$: $\mathbf{B}^{t-1} = \mathbf{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{d_1,d_2\}})\mathbf{V}^\top$
$\tilde{\sigma}_j = \mathrm{sign}(\sigma_j)(|\sigma_j| - \gamma\lambda_1)_+$, for $j = 1, \ldots, \min\{d_1, d_2\}$
$\mathbf{B}_1^t = \mathbf{U}\mathrm{diag}(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_{\min\{d_1,d_2\}})\mathbf{V}^\top$
$(\mathbf{B}_2^t)_{jk} = \mathrm{sign}((\mathbf{B}_1^t)_{jk})(|(\mathbf{B}_1^t)_{jk}| - \gamma\lambda_2)_+$, for $j = 1, \ldots, d_1$, $k = 1, \ldots, d_2$
$\mathbf{B}^t = \mathbf{B}_2^t - \gamma\nabla\ell(\mathbf{B}_2^t)$
**until** convergence criterion is met

### 2.2. Application to quadratic linear regression

We consider the regression model with interaction effects

$$Q_\tau\left(y|\mathbf{x}\right) = \xi_0 + \sum_{j=1}^{p} \xi_j x_j + \sum_{j,k=1}^{p} \beta_{jk} x_j x_k, \tag{2.3}$$

where $\mathbf{x} = (x_1, \ldots, x_p)^\top$ is the $p$-dimensional covariate, $\xi_0$ is the intercept, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)$ and $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{pp})$ are the main effects and interaction effects, respectively. For identifiability, we assume $\beta_{jk} = \beta_{kj}$. Model (2.3) can be expressed in matrix regression form by rearranging the coefficients into a matrix $\mathbf{B} \in \mathbb{R}^{(p+1)\times(p+1)}$, with $\mathbf{B}_{0,0} = \xi_0$, $\mathbf{B}_{j,0} = \mathbf{B}_{0,j} = \xi_j/2$, and $\mathbf{B}_{j,k} = \beta_{jk}$. In this way, model (2.3) becomes (2.1), with $\mathbf{Z} = (1, x_1, x_2, \ldots, x_p)^\top(1, x_1, x_2, \ldots, x_p)$. Dimension reduction in traditional interaction effects models often considers only the sparsity structure. The advantage of expressing the model in matrix form is that we can impose a sparsity assumption and a low-rankness assumption to further reduce the dimension, which is useful when the number of nonzero entries is still large.

Because $\mathbf{B}$ is a symmetric matrix, the estimate $\widehat{\mathbf{B}}$ should be the minimizer of the objective function (2.2) in the set of symmetric matrices in $\mathbb{R}^{(p+1)\times(p+1)}$. The incremental proximal gradient method can also deal with the case easily by changing the original gradient step by $\mathbf{B}^t = P_{\boldsymbol{S}^{p+1}}\left(\mathbf{B}_2^t - \gamma\nabla\ell(\mathbf{B}_2^t)\right)$, where $\boldsymbol{S}^{p+1}$ is the set of symmetric matrices in $\mathbb{R}^{(p+1)\times(p+1)}$ and $P_{\boldsymbol{S}^{p+1}}$ denotes the projection on $\boldsymbol{S}^{p+1}$. This can be written more explicitly as $\mathbf{B}^t = (1/2)[\big(\mathbf{B}_2^t - \gamma\nabla\ell(\mathbf{B}_2^t)\big)^\top +$

$\left(\mathbf{B}_2^t - \gamma\nabla\ell(\mathbf{B}_2^t)\right)]$. However, it is easy to see that as long as the initial value of $\mathbf{B}$ is symmetric, all subsequent steps still produce a symmetric matrix, and the projection step is redundant.

## 3. Theoretical Properties

In this section, we establish an upper bound for the estimation error of $\widehat{\mathbf{B}}$ obtained from (2.2) in a high-dimensional scenario. There are two key elements that allow us to derive the upper bound, following the pioneering work of Negahban et al. (2012). The first is the concept of decomposability for penalties. For a subspace $\mathbb{M} \subset \mathbb{R}^{d_1\times d_2}$, define its orthogonal complement as

$$\mathbb{M}^{\perp} = \left\{\mathbf{V} \in \mathbb{R}^{d_1\times d_2}; \langle \mathbf{U}, \mathbf{V}\rangle = 0 \text{ for all } \mathbf{U} \in \mathbb{M}\right\}.$$

Given a pair of subspaces $\mathbb{M} \subseteq \bar{\mathbb{M}} \subset \mathbb{R}^{d_1\times d_2}$, a regularizer $\mathcal{R}$ is decomposable with respect to $(\mathbb{M}, \bar{\mathbb{M}}^{\perp})$ if

$$\mathcal{R}(\mathbf{U}+\mathbf{V}) = \mathcal{R}(\mathbf{U}) + \mathcal{R}(\mathbf{V}), \text{ for all } \mathbf{U} \in \mathbb{M} \text{ and } \mathbf{V} \in \bar{\mathbb{M}}^{\perp}.$$

When $\mathbf{B}$ is a rank-$r$ matrix with $r \le \min\{d_1, d_2\}$, let $\mathbb{U} \subseteq \mathbb{R}^{d_1}$ and $\mathbb{V} \subseteq \mathbb{R}^{d_2}$ be a pair of $r$-dimensional subspaces spanned by the left and right singular vectors of $\mathbf{B}$, respectively. Consider the subspaces

$$\mathbb{M}_1 = \left\{\mathbf{A} \in \mathbb{R}^{d_1\times d_2} | \mathrm{row}(\mathbf{A}) \subseteq \mathbb{V}, \mathrm{col}(\mathbf{A}) \subseteq \mathbb{U}\right\},$$
$$\bar{\mathbb{M}}_1^{\perp} = \left\{\mathbf{A} \in \mathbb{R}^{d_1\times d_2} | \mathrm{row}(\mathbf{A}) \subseteq \mathbb{V}^{\perp}, \mathrm{col}(\mathbf{A}) \subseteq \mathbb{U}^{\perp}\right\},$$

where $\mathrm{row}(\mathbf{A})$ and $\mathrm{col}(\mathbf{A})$ are the row and column spaces, respectively, for the matrix $\mathbf{A}$. It is known that $\mathcal{R}_1$ is decomposable with respect to $(\mathbb{M}_1, \bar{\mathbb{M}}_1^{\perp})$. For the sparsity penalty $\mathcal{R}_2$, let $\mathcal{S} \subseteq \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$ be the indices of the nonzero entries with cardinality $|\mathcal{S}| = s$, and let $\mathcal{S}^{\perp} = \{1, \ldots, d_1\} \times \{1, \ldots, d_2\} \backslash \mathcal{S}$. Then, $\mathcal{R}_2$ is decomposable with respect to $(\mathbb{M}_2, \bar{\mathbb{M}}_2^{\perp})$, where

$$\mathbb{M}_2 = \bar{\mathbb{M}}_2 = \left\{\mathbf{A} \in \mathbb{R}^{d_1\times d_2} | \mathbf{A}_{ij} = 0 \text{ for all } (i,j) \in \mathcal{S}^{\perp}\right\},$$
$$\bar{\mathbb{M}}_2^{\perp} = \left\{\mathbf{A} \in \mathbb{R}^{d_1\times d_2} | \mathbf{A}_{ij} = 0 \text{ for all } (i,j) \in \mathcal{S}\right\}.$$

The second property concerns the restricted set that $\widehat{\mathbf{B}} - \mathbf{B}$ can be proved to be in. Let $\mathbf{P}_{\mathbf{U}^{\perp}}$ and $\mathbf{P}_{\mathbf{V}^{\perp}}$ be the projection matrices to spaces $\mathbb{U}^{\perp}$ and $\mathbb{V}^{\perp}$, respectively. Then, for a matrix $\mathbf{\Delta}$, define $\mathbf{\Delta}'' = \mathbf{P}_{\mathbf{U}^{\perp}}\mathbf{\Delta}\mathbf{P}_{\mathbf{V}^{\perp}} \in \bar{\mathbb{M}}_1^{\perp}$ (this is actually the projection of $\mathbf{\Delta}$ on $\bar{\mathbb{M}}_1^{\perp}$) and $\mathbf{\Delta}' = \mathbf{\Delta} - \mathbf{\Delta}'' \in \bar{\mathbb{M}}_1$. In addition, we denote by $\mathbf{\Delta}_{\mathcal{S}}$ the matrix in which $(\mathbf{\Delta}_{\mathcal{S}})_{ij} = \mathbf{\Delta}_{ij}$ if $(i,j) \in \mathcal{S}$, and $(\mathbf{\Delta}_{\mathcal{S}})_{ij} = 0$ if $(i,j) \notin \mathcal{S}$ ($\mathbf{\Delta}_{\mathcal{S}}$ is the projection of $\mathbf{\Delta}$ on $\mathbb{M}_2$). Then, the restricted set in our setting is defined as

$$\mathbb{C} = \left\{ \mathbf{\Delta} |\ \lambda_1 \mathcal{R}_1(\mathbf{\Delta}'') + \lambda_2 \mathcal{R}_2(\mathbf{\Delta}_{\mathcal{S}^\perp}) \le 3\lambda_1 \mathcal{R}_1(\mathbf{\Delta}') + 3\lambda_2 \mathcal{R}_2(\mathbf{\Delta}_{\mathcal{S}}). \right\}.$$

The value 3 in the above is somewhat arbitrary, and can be replaced by any constant larger than one. For convenience in the theoretical analysis, we write $\lambda_1 = \lambda\alpha, \lambda_2 = \lambda(1-\alpha)$, with $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \lambda_1/\lambda$. Then, the restricted set can also be written as

$$\alpha \mathcal{R}_1(\mathbf{\Delta}'') + (1-\alpha)\mathcal{R}_2(\mathbf{\Delta}_{\mathcal{S}^\perp}) \le 3(\alpha \mathcal{R}_1(\mathbf{\Delta}') + (1-\alpha)\mathcal{R}_2(\mathbf{\Delta}_{\mathcal{S}})).$$

Let $p = d_1 d_2$, $\mathbf{z}_i = \mathrm{vec}(\mathbf{Z}_i)$. In order to obtain the upper bound, we assume the following conditions. In the following, $C$ denotes a generic positive constant, the value of which can change between instances.

C1. $\mathbf{J} := E[\mathbf{z}_i \mathbf{z}_i^\top]$ is positive definite with its maximum eigenvalue $\sigma_{\max}(\mathbf{J})$ bounded by a constant.

C2. $\mathbf{z}_i = \mathrm{vec}(\mathbf{Z}_i)$ is sub-Gaussian in the sense that there exists a constant $C > 0$, such that for any unit norm vector $\mathbf{a}$, we have $E[e^{t\mathbf{a}^\top \mathbf{z}_i}] \le e^{Ct^2}$, for $\forall t > 0$.

C3. With $\mathbf{B}$ denoting the true coefficient matrix, there is a constant $c_1 > 0$ such that $E[\rho_\tau(y_i - \langle \mathbf{B} + \mathbf{\Delta}, \mathbf{Z}_i \rangle)] - E[\rho_\tau(y_i - \langle \mathbf{B}, \mathbf{Z}_i \rangle)] \ge c_1(\|\mathbf{\Delta}\|^2 \wedge \|\mathbf{\Delta}\|_F)$, for all $\mathbf{\Delta} \in \mathbb{C}$, where $\|.\|_F$ denotes the Frobenius norm.

Condition C1 is a mild moment assumption. The sub-Gaussianity of $\mathbf{Z}_i$ is required to bound different norms of a certain random matrix, as in Lemma 2 in the Supplement Material, and such a light-tail condition is often used in high-dimensional asymptotic analysis. Finally, C3 can be verified using more primitive assumptions, including the boundedness conditions for the conditional density of $y_i$ given $\mathbf{Z}_i$ and $\inf_{\mathbf{\Delta} \in \mathbb{C}} (E|\langle \mathbf{\Delta}, \mathbf{Z}_i \rangle|^2)^{3/2} / E|\langle \mathbf{\Delta}, \mathbf{Z}_i \rangle|^3 > 0$. Furthermore, the latter can be satisfied when, for example, $\mathbf{Z}_i$ is Gaussian. The proof is similar to that of Lemma 4 (3.7) of Belloni and Chernozhukov (2011), as shown in the Supplementary Material.

**Theorem 1.** *Suppose the true parameter* $\mathbf{B}$ *has rank* $r$ *and* $s$ *nonzero entries, and assumptions* C1–C3 *hold. If* $\alpha \in [0,1]$ *and* $\lambda \ge C \min\{\sqrt{(d_1+d_2)/n\alpha^2}, \sqrt{\log p/(n(1-\alpha)^2)}\}$ *for a sufficiently large* $C > 0$*, with probability approaching one, we have*

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \le C\lambda(\alpha\sqrt{r} + (1-\alpha)\sqrt{s}),$$

*as long as the right-hand size above is* $o(1)$*. In particular, taking* $\lambda \asymp C \min\{\sqrt{(d_1+d_2)/n\alpha^2}, \sqrt{\log p/(n(1-\alpha)^2)}\}$,

$$\|\widehat{\mathbf{B}} - \mathbf{B}\|_F \leq C \min\left\{ \sqrt{\frac{(d_1+d_2)r}{n}} + \frac{1-\alpha}{\alpha}\sqrt{\frac{s\log p}{n}}, \frac{\alpha}{1-\alpha}\sqrt{\frac{(d_1+d_2)r}{n}} + \sqrt{\frac{s\log p}{n}} \right\}.$$

Note that we allow $d_1$ and $d_2$ (and so does $p = d_1 d_2$) to diverge with $n$. On the other hand, the growth rates of $d_1$, $d_2$, and $s$ must satisfy $\lambda(\alpha\sqrt{r} + (1-\alpha)\sqrt{s}) = o(1)$. The theorem shows that the estimator can track the better performer of the nuclear-norm penalized estimator and the sparse (lasso) estimator. When $\alpha$ is sufficiently close to one, the rate becomes $\sqrt{(d_1+d_2)r/n}$, which is the same as the rate in Negahban and Wainwright (2011) for least squares regression. On the other hand, when $\alpha \approx 0$, the rate becomes $\sqrt{s \log p/n}$, as in Belloni and Chernozhukov (2011).

**Remark 1.** Theorem 1 establishes the error bound for a single quantile level $\tau \in (0,1)$. Suppose now model (2.1) is true for $\tau \in [\tau_L, \tau_U] \subset (0,1)$. When considering the uniform error bound for $\tau \in [\tau_L, \tau_U]$, an additional condition on the true coefficient matrix $\mathbf{B}(\tau)$ is needed. That is, there exist a (diverging) constant $L > 0$ such that

$$\|\mathbf{B}(\tau) - \mathbf{B}(\tau')\|_F \leq L|\tau - \tau'|, \text{ for all } \tau, \tau' \in [\tau_L, \tau_U].$$

Then, is we make assumption C3 also uniform over $\tau$, by following the same proof strategy as in Belloni et al. (2019), we expect to establish the same bound uniformly over $\tau \in [\tau_L, \tau_U]$. However, we leave the details out and focus on the single $\tau$ case here.

**Remark 2.** When $\alpha = 0$, the rate is only near oracle. We think that employing the adaptive lasso penalty $\sum_{j,k=1} w_{jk}|\mathbf{B}_{j,k}|$, where $w_{jk} = 1/|\tilde{\mathbf{B}}_{j,k}|$ and the initial estimator $\tilde{\mathbf{B}}$ is obtained using a lasso penalized quantile regression, would lead to the oracle rate, under additional conditions that involve a signal strength requirement, that is, a lower bound on $\inf_{(j,k)\in S} |\mathbf{B}_{j,k}|$. Signal strength conditions can be restrictive, but are usually required for the oracle property; see, for example, Zhao and Yu (2006), Meinshausen and Bühlmann (2006), Bühlmann and Van De Geer (2011), Zheng, Peng and He (2015), and Ndaoud (2019). A nonconvex penalty can also possibly achieve the oracle rate under such conditions. It would be interesting to establish the oracle rate for quantile matrix regression with both sparsity and low-rankness constraints, which is left to further research.

## 4. Numerical Results

We consider the quadratic quantile regression problem. The response is obtained using $y_i = \langle \mathbf{B}, \mathbf{Z}_i \rangle + \varepsilon_i$, where $\mathbf{Z}_i = (1, x_{i1}, \ldots, x_{ip})^\top (1, x_{i1}, \ldots, x_{ip})$, with $x_{ij}$ generated independently from a standard normal distribution, and the random error is generated as $\varepsilon_i = (1 + 0.2|x_{i,1}|)\epsilon_i$, with $\epsilon_i \sim N(-q_\tau, \sigma^2)$, where $q_\tau$ is the $\tau$th quantile of the Gaussian distribution $N(0, \sigma^2)$. The coefficient $\mathbf{B}$ is a rank-$r$ symmetric matrix obtained using $\mathbf{UDU}^\top$, where $\mathbf{U} \in \mathbb{R}^{(p+1)\times r}$ is the top $r$ left singular vectors of a matrix with independent standard normal entries. In order to generate a sparse matrix $\mathbf{B}$, we first generate $\mathbf{U} \in \mathbb{R}^{p' \times r}$, and then insert $p + 1 - p'$ zero rows into $\mathbf{U}$. Let $q = p'/(p + 1)$ be the proportion of zero rows in $\mathbf{U}$. We investigate the effect of different values of $q$ in our simulations.

Here, we apply the proposed method to estimate the coefficient matrix $\mathbf{B}$. First, we set the sample size $n = 300$, 500, and 700, and the dimension is set to $p = 30$. The true rank $r$ is 3, and we set $q = 0.5$ and $\sigma = 3$. The tuning parameters $\lambda_1$ and $\lambda_2$ are selected using five-fold cross-validation, and the step size $\gamma$ is always set to 0.1. We use $\|\widehat{\mathbf{B}} - \mathbf{B}\|_F$ as the errors reported in the simulation results. All simulations are repeated 200 times. Figure 1 compares our method with the lasso approach for the model with interactions. We see the errors decrease with $n$, and our approach outperforms the lasso as expected.

In the results reported in Figure 2, we set $n = 500$, and $q = 0.3$, and vary the dimension $p \in \{30, 50, 70\}$. In Figure 3, we report the results with $n = 500$, $p = 30$, and varying $q \in \{0.3, 0.5, 0.7, 0.9, 1\}$ (corresponding to about 8%, 23%, 46%, 76%, and 100%, respectively, nonzero entries). It can be seen that our approach outperforms the lasso in all cases, and the improvement is larger when $\mathbf{B}$ is denser.

Moreover, we compare the proposed quantile regression approach at $\tau = 0.5$ with the low-rank matrix mean regression Negahban and Wainwright (2011). For both mean and 0.5 quantile regression, we use $\|\mathbf{B}\|_*$, $\|\mathbf{B}\|_1$, or $\alpha\|\mathbf{B}\|_* + (1-\alpha)\|\mathbf{B}\|_1$ as regularizers. We take $n = 500$, $r = 3$, $q = 0.3$, and $p = 30$, 50 and 70, and the random error is generated from $N(0, 1)$ and $t(3)$. The results reported in Table 1 show that the performance of a mean regression may be better than that of a median regression when the random errors follow a standard normal distribution (but not always so, probably because we have heterogeneous errors here). However, a median regression outperforms a mean regression with heavy-tailed errors. The computing times of different methods are reported in Table 2. The settings are the same as those in the simulations.

Finally, we apply quadratic regression to nine regression problems from the

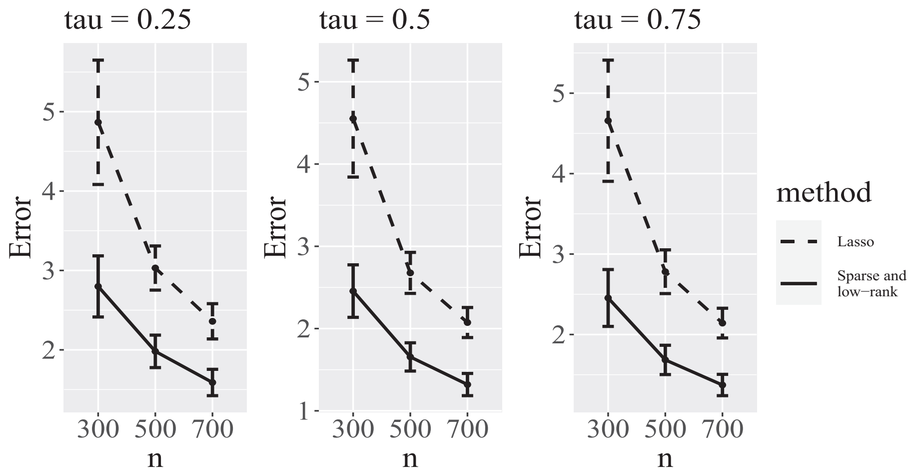


Figure 1. Estimation errors at quantile levels $\tau = 0.25, 0.5$, and $0.75$ when $p = 30$, $r = 3$, $q = 0.5$, and $\sigma = 3$. The error bars represent $\pm$ one standard deviation.

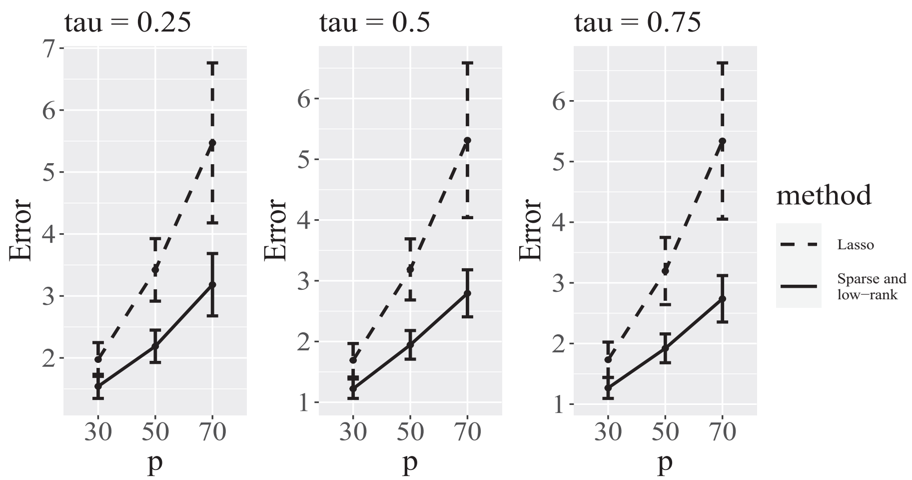


Figure 2. Estimation errors at quantile levels $\tau = 0.25, 0.5$ and $0.75$ when $n = 500$, $r = 3$, $q = 0.3$ and $\sigma = 3$. The error bars represent $\pm$ one standard deviation.

UCI machine learning repository. For each problem, we compare the proposed estimator with the lasso estimator (with interaction effects). The test errors are obtained using cross-validation, and the tuning parameters are chosen using five-fold cross-validation on the training set. The results are shown in Table 3. It can be seen that introducing the low-rank regularizer improves the performance.

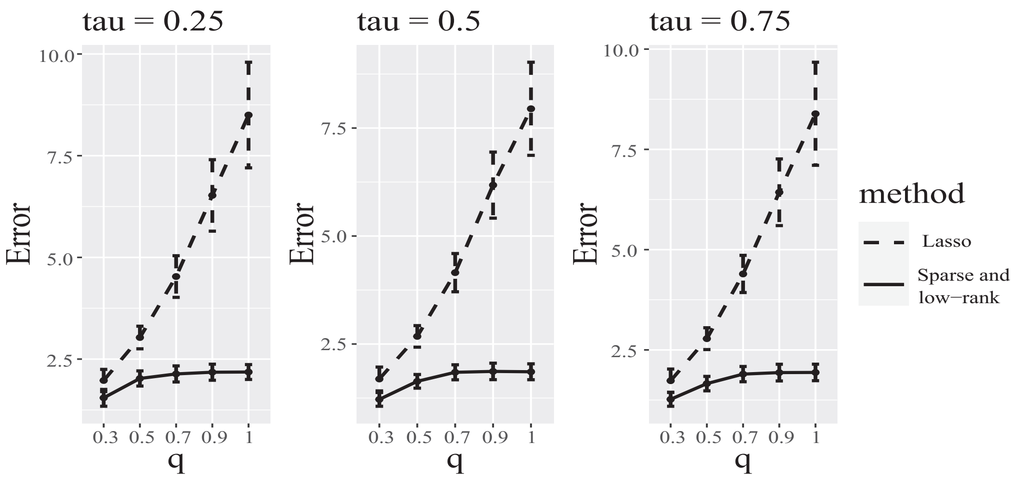


Figure 3. Estimation errors at quantile levels $\tau = 0.25, 0.5$, and 0.75 when $n = 500$, $p = 30$, $r = 3$, $\sigma = 3$, and $q$ ranges from 0.3 to 1. The error bars represent ± one standard deviation.

Table 1. Estimation errors for the proposed method (sparse and low rank) at quantile level $\tau = 0.5$ and mean regression (least square) when $n = 500$, $r = 3$, $q = 0.3$, and $p = 30$, 50, and 70. Numbers in parentheses denote the standard errors.

| | Regularizer | Method | $p = 30$ | $p = 50$ | $p = 70$ |
|---|---|---|---|---|---|
| $N(0,1)$ | $\|\mathbf{B}\|_*$ | 0.5 quantile | 0.60(0.08) | 1.21(0.19) | 2.72(0.61) |
| | | Mean | 0.62(0.09) | 1.81(0.38) | 3.28(0.68) |
| | $\|\mathbf{B}\|_1$ | 0.5 quantile | 0.58(0.10) | 1.81(0.47) | 4.83(1.45) |
| | | mean | 0.52(0.09) | 2.37(0.62) | 4.85(1.33) |
| | $\|\mathbf{B}\|_*$ and $\|\mathbf{B}\|_1$ | 0.5 quantile | 0.43(0.06) | 0.82(0.10) | 1.64(0.25) |
| | | mean | 0.38(0.05) | 0.74(0.14) | 1.55(0.34) |
| $t(3)$ | $\|\mathbf{B}\|_*$ | 0.5 quantile | 0.75(0.11) | 1.49(0.23) | 3.03(0.66) |
| | | Mean | 0.88(0.16) | 2.02(0.38) | 3.40(0.70) |
| | $\|\mathbf{B}\|_1$ | 0.5 quantile | 0.73(0.13) | 2.19(0.58) | 5.07(1.46) |
| | | mean | 0.88(0.21) | 2.59(0.62) | 4.98(1.32) |
| | $\|\mathbf{B}\|_*$ and $\|\mathbf{B}\|_1$ | 0.5 quantile | 0.52(0.07) | 1.00(0.13) | 1.88(0.32) |
| | | mean | 0.60(0.12) | 1.07(0.20) | 1.90(0.42) |

Table 2. Average computing times (in second) of the proposed sparse and low-rank method, lasso, and least squares approaches to complete the simulations, using R (version 3.6.3) on our desktop computer with a 3.40 GHz CPU.

| | $p = 30$ | $p = 50$ | $p = 70$ |
|---|---|---|---|
| Sparse and low-rank | 68.71 | 144.36 | 322.69 |
| Lasso | 24.58 | 36.81 | 113.46 |
| Least square | 86.38 | 168.65 | 236.29 |

Table 3. Test errors for nine regression problems at quantile levels $\tau = 0.25$, 0.5, and 0.75.

| dataset | $n$ | $p$ | $\tau$ | Lasso | Sparse & low-rank |
|---|---|---|---|---|---|
| Wisconsin Prognostic Breast Cancer | 155 | 32 | 0.25 | 1.46 | 0.93 |
| | | | 0.5 | 2.54 | 1.22 |
| | | | 0.75 | 1.89 | 0.81 |
| Residential Building–Sales Price | 298 | 26 | 0.25 | 0.18 | 0.11 |
| | | | 0.5 | 0.21 | 0.09 |
| | | | 0.75 | 0.08 | 0.07 |
| Residential Building–Construction | 298 | 26 | 0.25 | 0.21 | 0.14 |
| | | | 0.5 | 0.10 | 0.06 |
| | | | 0.75 | 0.07 | 0.04 |
| Real Estate Valuation | 331 | 6 | 0.25 | 2.34 | 2.03 |
| | | | 0.5 | 2.85 | 2.72 |
| | | | 0.75 | 2.55 | 2.37 |
| Forest Fires | 414 | 10 | 0.25 | 0.28 | 0.28 |
| | | | 0.5 | 0.58 | 0.53 |
| | | | 0.75 | 0.54 | 0.52 |
| Geographical Original of Music– Latitude | 847 | 68 | 0.25 | 1.91 | 0.97 |
| | | | 0.5 | 0.68 | 0.56 |
| | | | 0.75 | 1.23 | 0.44 |
| Geographical Original of Music– Longitude | 847 | 68 | 0.25 | 2.72 | 1.47 |
| | | | 0.5 | 1.31 | 1.14 |
| | | | 0.75 | 1.23 | 0.98 |
| PM2.5 Beijing–Aotizhongxin | 1,071 | 11 | 0.25 | 0.12 | 0.10 |
| | | | 0.5 | 0.12 | 0.10 |
| | | | 0.75 | 0.09 | 0.08 |
| Wine Quality–Red | 1,279 | 11 | 0.25 | 0.20 | 0.19 |
| | | | 0.5 | 0.26 | 0.24 |
| | | | 0.75 | 0.22 | 0.21 |

## 5. Conclusion

In this paper, we have proposed a convex regularized optimization approach for quantile regression with matrix covariates. The motivation for our work is the wide application of matrix regression and the lack of studies on matrix quantile regression. In order to reduce the effective number of parameters in the high-dimensional setting, two regularizers corresponding to low rankness and sparsity are imposed at the same time. We establish the upper bound on the estimation error of the proposed estimator and develop an algorithm based on the incremental proximal gradient. We apply the proposed method to quadratic quantile regression, where the covariates and their interactions can be reformed into a matrix. The advantage of the proposed method in quadratic regression problems is demonstrated using simulations and a real-data analysis.

When studying quadratic regression, the hierarchy restriction, that an interaction can only be included in the model if both or either main effects are selected, is often assumed; see, for example, Bien, Taylor and Tibshirani (2013), and Hao and Zhang (2014). When using the entry-wise lasso as a sparsity regularizer, a hierarchical structure is not incorporated. Strong heredity (an interaction effect can be selected only if both main effects are selected) can be incorporated by replacing $\|\mathbf{B}\|_1$ with a hierarchical penalty, for example, the composite absolute penalty in (Zhao, Rocha and Yu (2009))

$$\mathcal{R}_2(\mathbf{B}) = \sum_{j,k=1}^{p} \left(|\mathbf{B}_{j,k}| + \| \left(\mathbf{B}_{j,0}, \mathbf{B}_{0,k}, \mathbf{B}_{j,k}\right) \|_2\right).$$

The theoretical guarantee for this hierarchical penalty are left for further work.

## Supplementary Material

Proofs of the theorems are contained in the online Supplementary Material.

## Acknowledgments

We sincerely thank the editor, associate editor, and two anonymous reviewers for their insightful comments. The research of Zhongyi Zhu was supported by National Natural Science Foundation of China (11731011, 11690013, 12071087). The research of Heng Lian was supported by Project 11871411 from the NSFC and CityU Shenzhen Research Institute, and by Hong Kong General Research Fund 11301718, 11300519, and 11300721.

## References

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics* **40**, 1171–1197.

Argyriou, A., Evgeniou, T. and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning* **73**, 243–272.

Argyriou, A., Micchelli, C. A. and Pontil, M. (2010). On spectral learning. *Journal of Machine Learning Research* **11**, 935–953.

Belloni, A. and Chernozhukov, V. (2011). $\ell_1$-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* **39**, 82–130.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Fernández-Val, I. (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics* **213**, 4–29.

Bertsekas, D. P. (2011). Incremental proximal methods for large scale convex optimization. *Mathematical Programming* **129**, 163–195.

Bien, J., Taylor, J. and Tibshirani, R. (2013). A Lasso for hierarchical interactions. *The Annals of Statistics* **41**, 1111–1141.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.

Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics* **39**, 1282–1309.

Bunea, F., She, Y. and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics* **40**, 2359–2388.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE* **98**, 925–936.

Chao, S.-K., Volgushev, S. and Cheng, G. (2017). Quantile processes for semi and nonparametric regression. *Electronic Journal of Statistics* **11**, 3272–3331.

Chen, K., Chan, K.-S. and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 203–221.

Choi, N. H., Li, W. and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* **105**, 354–364.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**, 1348–1360.

Fan, Y., Kong, Y., Li, D. and Zheng, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics* **43**, 1243–1272.

Hao, N., Feng, Y. and Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association* **113**, 615–625.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **109**, 1285–1301.

Kato, K. (2011). Group Lasso for high dimensional sparse quantile regression models. *arXiv preprint arXiv:1103.1458.*

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

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

Koltchinskii, V. (2011). Von neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics* **39**, 2936–2973.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* **39**, 2302–2329.

Ma, Z., Ma, Z. and Sun, T. (2020). Adaptive estimation in two-way sparse reduced-rank regression. *Statistica Sinica* **30**, 2179–2201.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462.

Ndaoud, M. (2019). Interplay of minimax estimation and minimax support recovery under sparsity. In *Proceedings of Machine Learning Research* **98**, 1–22.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* **39**, 1069–1097.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research* **13**, 1665–1697.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. *Statistical Science* **27**, 538–557.

Pan, X. and Zhou, W.-X. (2020). Multiplier bootstrap for quantile regression: Non-asymptotic theory under random design. *Information and Inference: A Journal of the IMA* **10**, 813–861.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics* **39**, 887–930.

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

Yi, C. and Huang, J. (2017). Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics* **26**, 547–557.

Yu, L., Lin, N. and Wang, L. (2017). A parallel algorithm for large-scale nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics* **26**, 935–939.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 329–346.

Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37**, 3468–3497.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* **7**, 2541–2563.

Zheng, Q., Peng, L. and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *The Annals of Statistics* **43**, 2225–2258.

Zhou, H. and Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 463–483.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Wenqi Lu

School of Statistics and Data Science, Nankai University, Tianjin 300071, China.

E-mail: luwenqi@nankai.edu.cn

Zhongyi Zhu

School of Management, Fudan University, Shanghai 200433, China.

E-mail: zhuzy@fudan.edu.cn

Heng Lian

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong and CityU Shenzhen Research Institute, Shenzhen, China.

E-mail: henglian@cityu.edu.hk

(Received April 2021; accepted August 2021)